

Me, Myself, and My Hyperego: Understanding People Through the Aggregation of Their Digital Footprints

Aaron Robert Zinman

B.S., Cognitive Science
University of California, San Diego, June 2004

S.M., Media Arts and Sciences,
Massachusetts Institute of Technology, September 2006

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy in Media Arts and Sciences
at the Massachusetts Institute of Technology

September 2011

© Massachusetts Institute of Technology, 2011
All Rights Reserved

Author

Aaron Robert Zinman
Media Arts and Sciences
August 11th, 2011

Certified by
Pattie Maes

Associate Professor of Media Arts and Sciences
Alexander W. Dreyfoos, Jr. Chair
Thesis Supervisor

Accepted by
Mitchel Resnick

Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Me, Myself, and My Hyperego: Understanding People Through the Aggregation of Their Digital Footprints

Aaron Robert Zinman

B.S., Cognitive Science
University of California, San Diego, June 2004

S.M., Media Arts and Sciences,
Massachusetts Institute of Technology, September 2006

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy in Media Arts and Sciences
at the Massachusetts Institute of Technology

September 2011

Abstract

Every day, millions of people encounter strangers online. We read their medical advice, buy their products, and ask them out on dates. Yet our views of them are very limited; we see individual communication acts rather than the person(s) as a whole. This thesis contends that socially-focused machine learning and visualization of archived digital footprints can improve the capacity of social media to help form impressions of online strangers.

Four original designs are presented that each examine the social fabric of a different existing online world. The designs address unique perspectives on the problem of and opportunities offered by online impression formation. The first work, *Is Britney Spears Spam?*, examines a way of prototyping strangers on first contact by modeling their past behaviors across a social network. *Landscape of Words* identifies cultural and topical trends in large online publics. *Personas* is a data portrait that characterizes individuals by collating heterogeneous textual artifacts. The final design, *Defuse*, navigates and visualizes virtual crowds using metrics grounded in sociology. A reflection on these experimental endeavors is also presented, including a formalization of the problem and considerations for future research. A meta-critique by a panel of domain experts completes the discussion.

Thesis Supervisor: Pattie Maes
Associate Professor, Program in Media Arts and Sciences

Me, Myself, and My Hyperego: Understanding People Through the Aggregation of Their Digital Footprints

Aaron Robert Zinman

B.S., Cognitive Science
University of California, San Diego, June 2004

S.M., Media Arts and Sciences,
Massachusetts Institute of Technology, September 2006

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy in Media Arts and Sciences
at the Massachusetts Institute of Technology

September 2011

Doctoral Dissertation Committee

Advisor

Judith S. Donath

Fellow, Berkman Center for Internet and Society
Harvard University

Advisor

Pattie Maes

Associate Professor Media Arts and Sciences
Alexander W. Dreyfoos, Jr. Chair
Massachusetts Institute of Technology

Thesis Reader

Ryan Rifkin

Software Engineer
Google

For Bill Mitchell, Linda Jones, Rovine Jones, and Andrew Jones.
You are all sorely missed.

ACKNOWLEDGEMENTS

What a momentous occasion: to write the acknowledgements page at the every end of a giant seven year journey. This page is significant in that this thesis reflects those around me as much as it does myself. The Media Lab has to be one of the best places to “grow up” as a digital native. The constant exposure to its amazing community of tinkerers, doers, thinkers, builders, and crazies has illuminated how I can be a better me.

Thank you to everyone who has given me insight and pleasure. In particular:

Judith Donath, my mentor, for giving me enough freedom to learn by doing from your critique. Your breakthrough ideas will forever echo through me. Pattie Maes, my second mentor, for adopting me into your world and sharing your wisdom. Let the world be as elegantly practical as you (and always right)! Ryan Rifkin, my technical mentor, for helping me be a better engineer. Your balanced approach to work and life is refreshing and worthy of envy. And to the late Bill Mitchell,

My general exam committee members, Andrew McCallum and Jason Kaufman, for helping give me the structure for my vision.

My former mentors at IBM Research, Danny Sorokey and Chandra Narayanaswami, for asking the right questions to inspire another iteration.

My master’s committee members, Chris Schmandt and Walter Bender, for making me think a lot harder.

My undergraduate advisor, David Kirsh, who built my intellectual foundation. Very few at UCSD, let alone any university, are lucky enough to receive such a sophisticated pedagogy in such an amiable form. Marty Sereno whose genius is infectious and for letting me play beyond my capacity. John Batali for providing a unique perspective and supporting my work. Bill Griswold for his inspiration in ubiquitous computing that created my desire to come to the Media Lab.

The faculty of the Media Lab who directly helped shape my brain: Rosalind Picard, Chris Csikszentmihalyi, Henry Holtzman, Henry Lieberman, Sandy Pentland, and Deb Roy.

The miracle workers of the lab who do not get the credit they deserve: Linda Peterson, Aaron Solle, Jon Ferguson, Peter Planz, Will Glesnes, Jane Wojcik, Stacie Slotnick, Felice Gardner, Anna Carreiro, Deborah Widener, Paula Aguilera, Greg Tucker, Kevin Davis, Cornelle King, and Nicole Freedman.

Lab director Frank Moss for his instrumental help with Konbit, and new lab director Joichi Ito for exposing others to my work before he even officially started the job.

My evaluation committee, Ethan Zuckerman, danah boyd, Howard Rheingold, Con Kolivas, June Kinoshita, and Chris Marstall, for all of their deep insight. Their thoughts have altered the course of this work and beyond.

My UROPs over the years for all of their hard work, in particular Anh Nguyen, Erika Lee, and Jongmin Baek.

The Sociable Media Group for teaching me so much by being in your presence. I could not be prouder of my peers: Fernanda Viégas, Scott Golder, Christine Liu, Drew Harry, Dietmar Offenhuber, Francis Lam, Orkan Telhan, Yannick Assogba, and Alex Dragulescu.

The Fluid Interfaces Group for taking me in like your own, and teaching me a new perspective to my own work: Sajid Sadi, Doug Fritz, Marcelo Coelho, Seth Hunter, Pranav Mistry, Pol Pla i Conesa, Natan Linder, Roy Shilkrot, and Susanne Seitinger.

My family who have been the most supportive and loving possible. My positive qualities are directly a product of them: Mom & Dad, Elyse & Aaron & Maddy & Will, Grandpa & Grandma, Aunt Linda, Aunt Barbara, Aunt Lois, Uncle Byron & Diane & Joy & Isaias & Amber & Cobby, Ray & Marlene, David & Jill, Matthew & Colleen, Laurie, Jan Jannie Jan Jan & Don, Linda & Bob, Jack & Annelise, Helen & Asheesh, Eric & Sarah. Even those who have been lost will remain in our hearts and still have an impact on our day to day lives.

Anston Bosman who heroically helped get this document in order at the last minute.

My long time friends who have been my inspiration, cheerleaders, motivators, supporters, challengers, and benchmarks. I am so lucky to have such amazing friends. They are like red wine: they keep getting better with age. Erik Sparks, Greg Elliott, Eric Lau, Eric Wahl, Karuna Murdaya et al., Dan Peirano, Diana Ziola, Rhona Stuart, David Kim, Eric Nguyen, Bill Huey, Lisa Han, Gena Peditto, Roger Torino, Veronika Stelmakh, Lily Pollans, Rana Amir, Minna Ha, Jonah Knobler, Geoff Rawle, Stephen Miran, Robert Brown, Rohit Reddy, Howard Yu, David Francis, Aaron Glassenberg, Edwin Powers, Chee Xu, Suelin Chen, Aaron Tang, Paulina & Tom Süderland, Chris Collins, Jimmy Jia, Hugo Liu, Maya Orbach & Nadav Aharony, Daniel Olguin Olguin, Mirja Leinss, Olof Mathé, and Thomas Rebotier.

And all my friends at MIT and beyond: emax, Ali Mohammad, Amanda Parks, Anmol Madan, Noah Vawter, Dustin Smith, Bo Morgan, the entire Dreamteam, Cory Kidd, Jon Gips, Angela Chang, David Broniatowski, Jim Barabas, Jamie Zigelbaum, Jay Silver, Mako, Jeevan Kalanithi, David Merrill, Joost Bonsen, Jarrie Karahalios, Kwan Lee, Leo Bonanni, Luis Blackaller, Manas Mittal, Marcel Botha, Matt Hockenberry, Todd Farrell, Nathan Eagle, Neha Narula, Phil Liang, Nick Knouf, Pol Ypodimatopoulos, Rachel Kern, Richard The, Seth Raphael, Ben Waber, and Taemie Kim.

TABLE OF CONTENTS

1.	Introduction	13
2.	Background	27
3.	Related Work in Computer Mediated Communications	38
4.	Experiments	60
5.	Abstraction Techniques	129
6.	Reflections	139
7.	Evaluation.....	143
8.	Conclusion	157
9.	Bibliography	160
10.	Appendix A: Survey Results	173

1. INTRODUCTION

“You’re born naked and the rest is drag.”
--RuPaul

Dealing with strangers is fundamental to everyday urban life. We complete transactions with the bus driver and the barista, we help an elderly person up from their seat, and we make small talk at happy hour. In each interaction we assess unknown people by comparing our perceptions against our preexisting models of the world. These impressions are used to bias how we interpret the information we glean and help us decide whether and how we wish to continue our interaction. Perceiving and reacting to unknown people also occurs daily on the Internet. We read through strangers’ product reviews on Amazon, reviews of surgeons on Yelp, and profiles of those who wish to friend us on Facebook. We discover new DJs on Turntable.fm, follow coders on Github, and participate in Middle East revolutions via Twitter. We might even micro-finance grocery store owners in Peru using Kiva. However, our ability to perceive and interpret another person’s actions and intentions is far more limited online than in person. The cues available to us are typically isolated to a single brief textual communications act. Much of the nuance that comes from noticing someone’s clothing, gait, or even accent is not possible when cast in a limited virtualized form. We have to instead reinvent the rules and interpersonal strategies for each new mediated channel. Yet we are not entirely at a loss; we can enhance our understanding of a stranger by examining the content of their past acts juxtaposed against their situational context. As actions speak louder than words, there is much to be learned by understanding the larger behavioral trends of an individual or a collective. This thesis hypothesizes that socially-focused machine learning can help us form better impressions of online strangers through the analysis and display of their digital footprints.

The past decade has seen an explosion of novel Computer-Mediated Communications (CMC) technology. The euphoric rise of social networking and mobile devices has exponentially increased the participation and volume of information about ordinary people online by providing more ways to communicate with their friends. While this has been extraordinarily useful, the current interfaces for social networking are not well suited for strangers. The standard paradigm is to view a person in terms of their self-description and a reverse chronological ordering of their past actions usually from that single site. This works well enough for friends who already know each other and simply want the latest gossip, but this lens is a pinhole compared to what is needed for a stranger. Self-descriptions do provide some latent insights, but they are inherently unreliable and likely do not tell you what you really want to know (Hancock, Toma, & Ellison, 2007). A list of thousands of items is costly to traverse, does not differentiate its items, is of mixed

value and removed from its context, and presents a perspective that is at once biased and difficult to read. We need a different way for technology to help us better understand strangers online.

The need to form a deeper impression of online strangers is becoming increasingly urgent. Every day, millions of people online must decide whom they should trust when purchasing goods, which opinions/arguments are persuasive, who has credibility, and whom to date romantically or partner with in business. While some people who have a higher risk tolerance already happily engage with others online, there remain risk-adverse individuals for whom a given engagement is off-limits given the current interfaces. There are many benefits to improving such individuals' ability to gain a sophisticated impression of an online stranger. Key insights into a stranger's character help the observer with the necessary trust and judgement calls that are instrumental to interaction. With increases in engagement comes advancements in commerce, collaboration, socialization, and scenarios not yet imaginable.

It is important to solve issues surrounding online impression formation as much for individuals as it is for entire crowds of people. When individuals are aggregated, we can gain insights into our society that have real democratic consequences, as the social media-powered Arab Spring has recently demonstrated to the world. Just a few weeks before this writing, President Obama lead a virtual Town Hall meeting in which he took questions from ordinary citizens using Twitter. Aside from being a milestone for CMC, it represented an opportunity not just to respond to a few questions but to react to the stream as a whole. If deep insights could be compiled about each individual, such as their viewpoints and demographics, we could cluster around their concerns to gain a far richer and deeper impression of society's pulse than any New York Times survey could bring. Twitter is just one of many low-cost ways that can increase online engagement, in this case political, should the outcome give voice and attribution to those who Tweet. There are non-political opportunities that arise from the meaningful aggregation of individuals as well, such as letting job seekers preview company culture, assessing the team-spirit in an open source community, or setting expectations in joining a potential social club.

Powerful pseudonymity is another missed opportunity for a world that only focuses on one communications act at a time. If a top expert wishes to post a comment on a relevant NYTimes.com article, he or she may be drowned out in thousands of other comments. Depending on the stakes and provocation of their opinion, when read such an expert is likely to be doubted, cast as an impostor, or ignored. Instead of such an unfortunate fate, the comment could have more weight through the empirical backing of a history of germane data that is legible at a glance. Technology could help credibility within the online world be even more portable and evolving than it is for offline reputation. The data need not reveal specific identities, only showing enough partial evidence of activity or consistency in character as needed to prove a point without risking being unmasked. There are commercial incentives too: companies could

create better tailored ads and experiences by computing over large data bodies without needing to know personal details about an individual user¹.

In theory it should be possible to synthesize meaningful insights about an online persona that accurately reflect the offline person. We form impressions of others through their choices; prototypes are formed in part through examining choices of word, fashion, and taste. The same examined mind makes choices with a similar logic online as offline, which will be evident in what that person clicks and types. Over time, the aggregation of one's online behavior should converge on a reasonable approximation of who one is², or at least who one is within a given context. Helpfully, this data is becoming more abundantly available as our lives are increasingly networked, recorded, and broadcast.

This was not always possible. By default data remains in isolated database silos, imprisoned by proprietary schemas, access controls, and ideas of privacy and intellectual property. As it is extra work to free the data in a usable fashion, there would need to be a demand for adequate use cases before any software engineer would deem it worth the additional effort. The cultural conditions in industry are radically changing, both now and only more so in the future. These data silos are being freed with trends like virtualized or “cloud”-based operations, public APIs, OAuth which enables sites to share data with user consent, structured data, common standards, and purposely accessible and indexable information that draws traffic.

Previous attempts at visualizing social spaces have inherently relied upon direct mappings between the data set and the visual domain; in contrast, this thesis proposes to use *machine learning* techniques in order to push past the limitations of basic statistics to abstract and synthesize meaning. Machine learning is a phrase used to describe the ability for machines to recognize patterns in data, among other intelligent capabilities. While presenting any subset configuration of the available information about a person constitutes a bias, related work has attempted to stay more objective by focusing on visualizing structural details in a 1:1 mapping. Here, the term *structural* is used to refer to the variety of ways a communications technology is used without examining the content. For example, “who sends whom messages with what frequency” constitute structural measurements that can be communicated to the user straightforwardly. Structural details are useful and can help answer questions such as which members are the principal contributors to a given community (Gleave, Welser, Lento, & Smith, 2009), who might be a spammer (Zinman & Donath, 2007), and provide measurements of overall activity or

¹ This has been shown to be a tricky proposition, as many attempts to anonymize data have failed in the face of those motivated enough to discover the true identities (Barbaro & Zeller, 2006; Singh & Zhan, 2007).

² While there is no one “true” self, each self-presentation is a true self for that moment in time (Goffman, 1959).

liveliness (Xiong & Donath, 1999). These are useful in gaining an impression of an individual or crowd, but eventually they fall short of the deeper questions that could be asked.

This thesis focuses on approaches that abstract data from structure and content using machine learning to form more powerful impressions and insights. It presents four original ways in which persistent histories can be used to better understand strangers. The experiments address issues of heterogeneity of data type and source, scale, the crowd versus the individual, and structural versus semantic behaviors. Each experiment uses a different method to reflect a range of goals and audience, employing statistics, algorithms, and visualization to the cause of yielding the greatest value for all players involved.

1.1 Empirical history as opportunity and viewpoint

Current CMC designs present a dearth of cues about a person. We have a holdover from face-to-face (FtF) communication where each communications act is the main focus, as opposed to an approach that unifies the present equally with the past. Mediated communication first started in written or semaphore form, and stayed similar through the next set of channels: the telegram, the telephone, email, and online forums. Each new channel has remained focused on one act at a time, even in the Facebook newsfeed which is a radical departure from the past in other ways. Most electronic media present a name for the individual, the time that the message was created, and perhaps a few statistics about the person. In Figure 1.1 we can see that phpBB, a very popular open source package for forum discussions, presents the join date and number of posts of its members. We do not, however, see if they usually start discussions, reply to others, or are viewed favorably by the community. Nor do we know in which types of discussions they are most likely to engage, whether they have strong interpersonal skills, or have any expertise outside of what is contained in the message. Yet of all of this data may be available in the forum's archives; it just needs to be surfaced.

Yet surfacing this data is more complicated than simply making it available in a deep link. Many CMC interfaces allow you to traverse the history of an individual. Reading the past few comments or instances can be illuminating by itself. However, knowing which data or abstractions of data to highlight is a tricky problem because there are so many ways to slice and subset the data, and each one may be appropriate for some situations but not others. We do not, for instance, need to know the political ideologies of a stranger from whom we are purchasing a used motorcycle. However, if they were diligent in their political activism, that mentality could signal attention to detail and thus indicate thoughtfulness in past bike maintenance. There are, to be sure, nearly infinite other goals a user might have for another online. Every day, millions of people might want to know if a solicitor is credible enough to satisfy the requested followup, or if



Figure 1.1. A message on the tonymacx86.com forum running the popular phpBB discussion software. phpBB automatically presents the join date and number of posts from an individual, as well as a variety of other facts relevant to the enthusiast community.

the person in the dating profile would return their witty humor with sufficient banter. Because there are infinite goals a user could have and a different type of source data is needed for each goal, there can be no universal solutions to facilitate online impression formation.

With any CMC work, whether a tool for analyzing the past or a new medium to connect individuals in the present, there are numerous facets that a designer may choose to emphasize or deemphasize. Corporate instant messengers often integrate with company-wide computer authentication to ensure that the interlocutor will take responsibility for their actions by making their identity canonical. Political activists may require Tor-like systems (Dingledine et al., 2004) to mask their identity and history. There is no universal set of guidelines in the design of CMC because each mediated medium makes different trade-offs to match the problem domain. As such, there is not a catholic method to understand the spaces therein even under a specific user goal, because the way each channel is used is always context-specific.

It is the responsibility of the designer to consider the relevant factors for both medium design as well as any tools to aggregate and summarize the space. While it is unlikely that future usage will perfectly reflect the original assumptions, careful consideration is required as the affordances of a medium directly influence convention (Norman, 1998). For example, prior knowledge of the types of discussions can influence choice of interface or visualization (Dave et al., 2004). Preece (2001) notes that “*broad shallow threads are characteristic of empathic discussions whereas narrow deep threads are typically generated in discussions of factual information.*” Such knowledge can inform the structural representation by emphasizing either individual messages (e.g. mutt) or conversations (e.g. Gmail), which can then inform how we begin to aggregate an individual against the contexts they participated in.

Empirical behavior data within an online context gives us cues about a person. As it is not possible to read through all of everyone of interest's history, we must find ways to condense these histories into an intelligible gestalt. There is a variety of design and algorithmic approaches to do so, and this thesis later discusses some of those techniques in length. The past interactions we visualize need not be isolated from one site or another, but may be combined to give richer insights into an individual. Each site has its own context or use case, and through their combinations we gain more insights into an individual. We can see action and reaction, viewpoint and counterpoint, active versus passive usages, adoption or rejection of trends, spikes of concentrated behavior, preferences towards other classes of persons, and others' opinions.

Communicating those aspects in a way that is fair to the author, quickly understood, computationally tractable, and that helps the user answer the questions they are likely to have are an extraordinarily difficult set of tasks. We must recognize that those difficulties translate into a set of choices made by the designer. To emphasize the role of the designer and artist in how we computationally gain an impression of individuals and crowds, we use the term *data portraiture* to describe the end-result.

1.2. Data portraiture

In cyberspace, we are bodiless. Despite the obvious and long-desired advantages of removing race, gender, age, and other non-mental attributes from online interactions (Hiltz & Turoff, 1978), the physical body remains a powerful force in face-to-face interactions. Stereotyping allows society to function as a whole (Simmel, 1910), and minute physical gestures are important for efficient communication (Zebrowitz, 1997), trustworthiness (Handy, 1995), and expression of identity (Donath, 2007). In the art world, portraiture has a rich and venerable tradition that exploits our ability to recognize these physical properties to obtain a multidimensional gestalt of character, form, and function (Brilliant, 1991). Carrying over this tradition into the digital realm can help individuals not only make better sense of strangers in the online spaces they inhabit, but can also help organizations to understand their information flow, facilitate better collaboration, and function egocentrically as a digital mirror to better understand ourselves (Donath, 2010).

In post-Renaissance western Europe, portraiture was reserved as a way for the rich and powerful to encapsulate their accomplishments and status. Men would be painted with their weapons, symbolic or real. Noblemen differentiated themselves through clothing, stance, and scene. The meaning derived from a work become a mixture of projection of the subject through the lens of the artist. To be sure, artists have the benefit of human reasoning and expressive capacities to cast the subject in the light of their choosing. They can add emblematic objects to a scene, alter expression on the micro-level, and even change the light and colors to reflect a desired mood. Like machine learning techniques, artists carefully perform semantic compression of their subjects.

In data-driven portraiture, by contract, we do not always have the luxury of human intervention with each generated portrait. Nor should we; we gain the digital advantage of presenting extracted meaning from data on a large scale. None the less, much as the artist injects subjectivity into a portrait with every brush stroke, the data artist does the same. In selecting the choice of algorithm, data sets, stop words, and eventual visual representation, the data miner injects their choices into a domain that is often viewed as authoritative. We must take special care to ensure that the resulting presentation reflects the same amount of ambiguity inherent in the compression. This is difficult in the abstracted domain because we cannot easily rely on the preexisting categories and stereotypes that we normally use to infer unknown attributes of others (Simmel, 1910). Instead, we must pay careful attention to the tools of the abstract domain; visually this is often color, shape, and typography. Color effects, metaphors from the physical world, and cultural traditions can unexpectedly assign and alter meaning to different parameterizations of these abstract classes. For example, many cultures associate the color red with danger, violence, and passion -- but other cultures do not. Seemingly arbitrary choices like the hue range in a color spectrum can also affect interpretation of purely scientific data, despite the common brightness and saturation levels (Rogowitz & Treinish, 1996). See Rogowitz and Treinish (1996) for a useful discussion of issues and guidelines in visualizing scientific data.

Bearing these caveats in mind, this thesis argues that carefully designed data portraits can enable and facilitate exciting new applications. In a world increasingly overflowing with information, reliable methods of presenting raw and aggregated data have become urgently necessary. We believe that machine learning holds the power to push data portraiture from its 1:1 confines towards useful abstraction.

1.3. The problem of online impression formation

To expand the role of online impression formation and its subproblems, a formalized description must first be articulated. We break down the problem from a systems perspective to establish a working vocabulary and mental model for future researchers and data portraiture artists. The model is a representation of the problem in its most general form, separating out the subcomponents and their interactions. The formalization is followed by a set of questions and considerations that should be considered by any designer attempting to depict strangers online.

We must first define the problem of online impression formation to understand how to deconstruct and solve it. At first, the problem seems bounded and thus simple. There are three main players: the *subject* whose data is under examination, the *data modeler* or *artist* who is transforming that data, and the *observer* who is interpreting the transformed result to gain an impression of the *subject* to whatever end. The observer's goals may include judging the character of the subject to assess risk in commerce or offline personal safety, assessing their expertise,

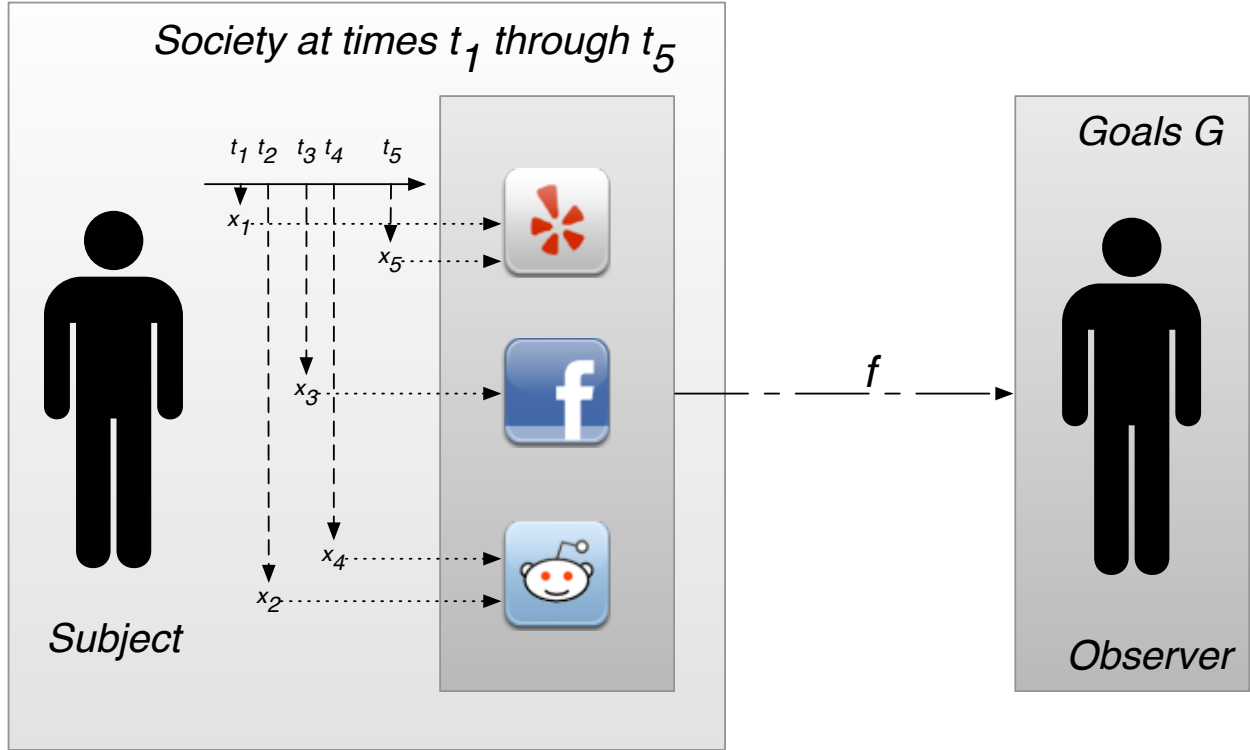


Figure 1.2. A subject performs communications acts x_1 - x_5 at times t_1 - t_5 across various online services. An observer may gain an impression of the subject based upon the presentation of data using transformation function f . The function should reflect the observer's goals, minimize the cost of acquiring the impression, and recognize the tension between objectivity and data compression.

predicting future social dynamics, and satisfying a general curiosity. Given a finite (but potentially large) amount of data available about a subject, there exists a *maximum impression* that could be yielded by an observer reading, listening, and watching each available datum. The impression is a complex mental representation involving a calculus of prototypes, contextually relevant issues, judgement and predictive models of choices, and sentiment. The amount of time to do so is the *total cost of perception*, yielding an upper bound. The problem of representation for a *data scientist* is to create a transformation function f in both visual and data domains that best approximates the maximum impression while minimizing the cost of perception. The problem of representation for a *data portraiture artist* is for f to maximize the impression possible through their distorted lens, focusing on certain details while omitting others. The data portraiture artist maybe strive for objectivity, as was the case for three out of four experiments described in this thesis. The difference lies in how they recognize the importance, sensitivities, and replaceability of those dimensions they select for f , usually those that demonstrate socially useful artifacts and prototypes.

While it is not clear how yet to approach f , this definition at first seems reasonable. People interact within sites, we have some ability to look at all their past history, and we wish to gain an

impression of them, subjective or objective. The problem lies in that our impressions are not stable or deterministic. The main complications are of contextual alignment and goal.

The subject's data was not created to be easily perceivable in some future alternative form, and as such is not self-explanatory. Each datum is a result of an isolated communications act within a single context with an intended audience at a specific time. The audience at time t for communication x is likely to change with future observers. As such, the context and its common understanding across the subject and their original audience are either missing or understood to be different. Therefore future observers will have a different and perhaps mischaracterized notion of x . Thus reveals the most challenging aspect: no act x contains all the necessary information to properly understand it as the author intends. It is always grounded within a culture and society (Clark & Brennan, 1991; Clark, 1996). Humans perform lossy compression of ideas to be able to communicate with a reasonable minimum of effectiveness. Machines are not yet good at decompressing such messages, which would require a simultaneous have command of language, socialization, embodied common sense, common ground, empathy, expression, and everything else it means to understand humans. Both objective and subjective representations benefit from more accurate computational analysis as it simply provides a better starting point.

We must recognize that in this formalization the data portraiture artist and objective data scientists are the ones that create the data transformation function f without the input of the subject. An alternative design may consider how subjects (or even observers) could influence f to give them control over how subjects are depicted, perhaps annotating errors, inaccuracies, and missing information as well as changes in personality or life predicaments over time. Control over self-presentation is a fundamental aspect of the existing social world and to ignore that is somewhat unnatural. This thesis takes the approach of exploring the boundaries of data mining and visualizing existing social data to scaffold future discussions surrounding online impression formation. Future works are in a better position to incorporate subjects' control and self-descriptions once the technological possibilities have been first explored.

GOAL-DRIVEN IMPRESSION FORMATION

When we seek to gain an impression of others, we often do so to accomplish a specific goal. Observers may make a variety of impressions, but ultimately those impressions will be biased towards any task at hand. For example, take the practical task of finding suitable baby sitters through online profiles. We wish to gain impressions of a variety of character traits that belong to a "good" babysitter. In recognizing that desirable babysitters are responsible, each babysitter will try to position themselves as responsible. For instance, a babysitter in an online profile may provide an anecdote about how they remained calm while calling poison control. A skeptical observer may call into question why poison control was needed in the first place if the babysitter

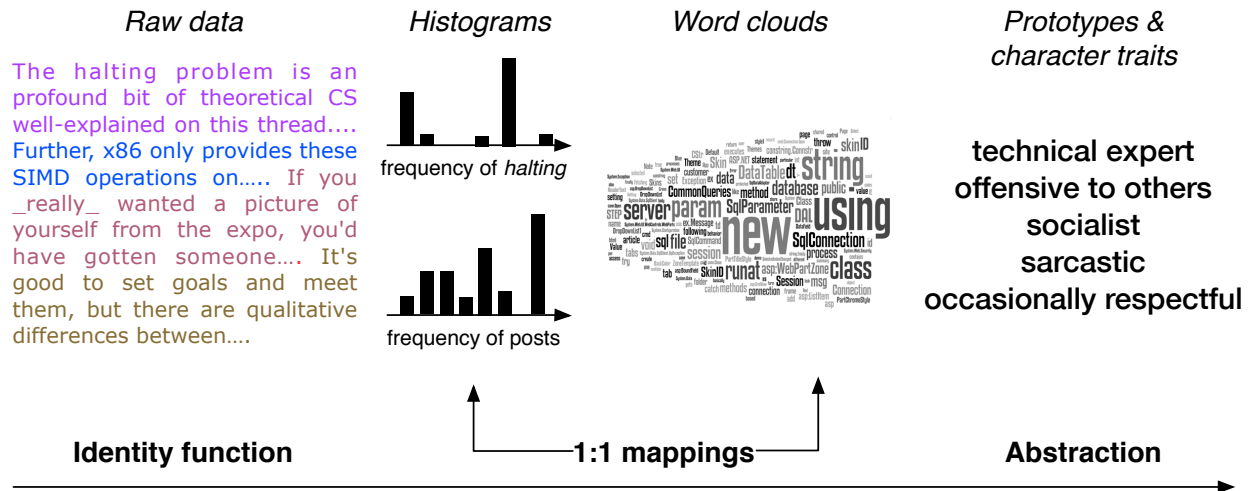


Figure 1.3. The axis of abstraction in transforming online data into a data portrait. We move from raw data towards 1:1 mappings that demonstrate basic statistics about the data, such as the frequency one uses each word. As we continue towards abstraction, we combine and aggregate the data with external models of the world to create new representations that likely will not be found verbatim in the content.

was in fact responsible. This example helps illustrate the complexities of representation given a goal: even humans may not be able to choose a proper representation to fit a goal, and the types of analysis and metrics undergone to assess the goal vary across individuals and situations.

There are three main possible function types to describe the data transformation function f . They all lie along the same axis of abstraction, as shown in Figure 1.3. On one extreme is simply showing the raw data, which would in mathematical terminology be called the *identity function*. Because the identity function is the most costly to perceive, unless the data is especially sparse it is not useful when some level of abstraction could helpfully be performed. The degree to which the data is abstracted either computationally or visually directly affects the subjectivity versus objectivity of f . Objectivity is better achieved by mapping from the data to the visual domain in a 1:1 fashion, such as with a histogram. Obviously which histograms are used is the subjective call of the data artist, but the underlying mapped data remains objectively plotted as it is unaltered. When possible, 1:1 mappings should be used because they are the easiest for users to understand how the visualization related to the raw data. On the other extreme lies true abstraction in both visual and data domains. Abstraction is the result of synthesis in the data domain to condense a large volume of information into a set of generalizations. Abstraction is necessary to achieve higher-level semantic summaries such as labeling individuals as a *technical expert*, a conclusion that comes from seeing trends and assessing expertise rather than histograms or other statistical plots.

Regardless of function, the perceivable output will be evaluated by the observer using background models of society, community, language and culture according to various goals. The chosen goals also impact impression formation, such that certain data patterns may be more

sought than others. Therefore the goal of a universally deterministic impression is impossible for a single observer and their goals, not to mention for a more general public. More likely, data artists will use the domain knowledge about each site to customize their designs around the existing demographics, culture, and expected goals.

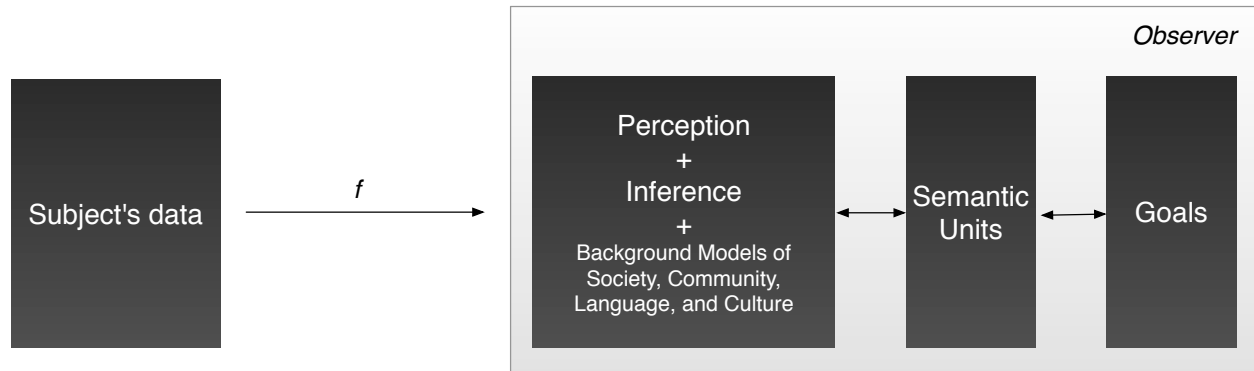


Figure 1.4. An observer attempts to match their goals to information provided by the subject. The data is transformed to the visual domain within three extremes of f : (a) An identity transformation returns the data in its native form, (b) The data is linearly mapped to the corresponding visual domain in a 1:1 fashion, or (c) The data is transformed by another function g , abstracting the data so as to subset and thus synthesize the mapped data.

ABSTRACTION

While the data may be mapped in a variety of forms, it is easiest for the observer if it can map directly to the semantic units that can satisfy a goal as shown in Figure 1.4. In the baby sitter example above, there are many intermediate representations that help clarify the goal of finding a good babysitter. For some class of observers, misspellings and poor grammar might be abstracted into the semantic units of “*poorly educated*” or “*lazy*.” As many observable data could equally conjure these semantic units, any representation that is not simply a tag that literally says “*poorly educated*” or “*lazy*” requires more effort from the user during impression formation than desired. Some effort may be required of observers to push the data portrait closer towards objectivity for ethical reasons given any inaccuracies and unnecessary bias.

Abstraction is most helpful if we can provide the same semantic units as would be perceived from the identity function. As the individually perceived semantic units cannot be known *a priori*, designers who choose to abstract the data must be cognizant of what they choose to leave in and leave out. Even though machines could classify résumés against a subjective abstraction of education quality, there are many more textures contained in the résumé leading to a host of different observer-performed characterizations. What other semantic units are discarded may be predicted by the observer using their background societal models, often incorrectly.

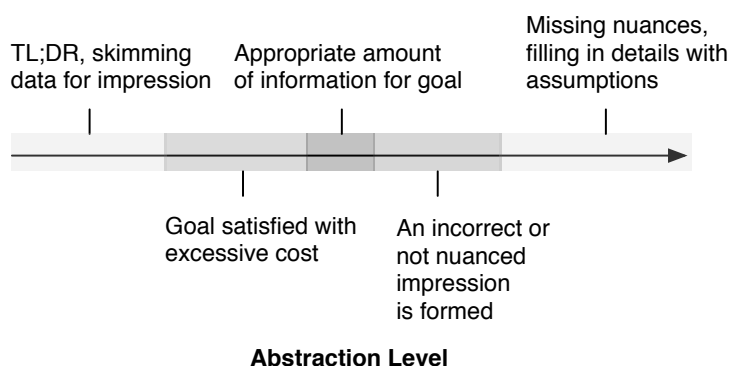
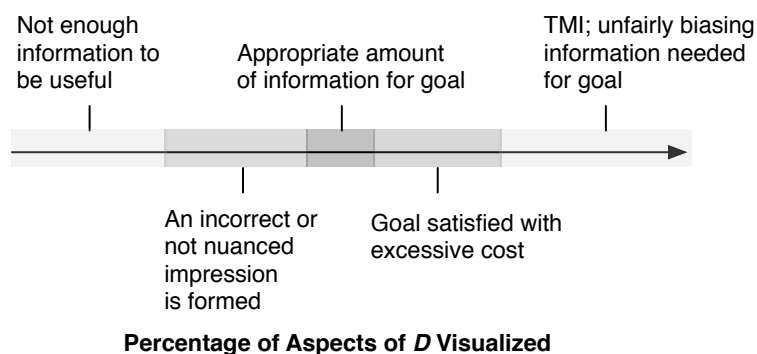


Figure 1.5. Continuums of information exposure and abstraction in online impression formation.

balance. *Defuse*, a project later described in the experiments chapter, started with a bias towards exposing the raw underlying data as to remain objective. Ultimately this led to too much data being presented, with the results that the user would not know where to look and would skip over the majority of what was present. Over time, shifts were made towards using a sociologically-driven structure to increase the abstraction. The new interface could better tell the users what the machine had identified using targeted filters that were built on premises of what the user might already want to know. Abstracting the data based upon the perceived goals of the users facilitated a more straightforward presentation.

Thus we can see a continuum of both abstraction and aspects revealed, where each have undesirable extremes as shown in Figure 1.5. Where we place a given visualization on both continuums has a direct interaction with its objectivity and subjectivity. While certain interaction strategies can allow individuals to dive deeper into data and thus help blur the distinction, ultimately more tools are needed for users to build their own models and filters so they are responsible for more of the subjectivity. This is particularly needed to better answer the “so

That said, however, leaving out other high level characterizations is not always a problem, and is indeed sometimes necessary. As a subject’s dataset grows large and spans an ever greater timeline, qualities will likely be revealed that are just as inappropriate as appropriate for a given set of goals, yet can still bias the impression even if irrelevant. For example, political ideology and religion are typically not germane for many contexts of interaction, but if known could alter further interaction by needlessly triggering prejudice. The data portraiture artist must make critical decisions about the level and amount of information revealed.

These issues can be addressed by an artist at the beginning of a data portrait, but likely it will take several iterations to find the correct

what?” question of understanding themes that may exist within a community. Unless we can dive deeper into the abstracted data and its connected trends, we may not know why certain semantic units may be less or more important.

1.4 Dissertation roadmap

The structure of the thesis is as follows:

Chapter 2 reviews the existing literature in impression formation in the FtF and CMC worlds. This allows us to think about how society already functions given the need to interact with strangers, and motivates ways we can think about processing online data. In particular, it outlines insights from the sociologists Ervin Goffman, Pierre Bourdieu, and Georg Simmel each of whom gives us guidance on the different social geometries and strategies that exist. A discussion of issues present in conveying online strangers follows.

Chapter 3 discusses related work within CMC. It examines past trends of social visualization, grouped into four main themes: 1) discovering the main players of a community, 2) monitoring the social health of a community, 3) uncovering relationships between individuals, and 4) diving into the semantics of past interactions. It concludes with a discussion of relevant commercial ventures.

Chapter 4 cover the novel experiments created for this thesis. The first, called *Is Britney Spears Spam?*, examines the structural and network activity of MySpace users to prototype them in social and promotional intention. Next, *Landscape of Words* builds a model of the topics discussed on Twitter, and uses it in turn to visualize the active topics of individuals, their surrounding networks, and Twitter as a whole. Following, *Personas* attempts to show users how the Internet sees them by visualizing the process of machine learning categorization of statements about a given name. Finally, *Defuse* provides an alternative interface for viewing comment and commenters online, using each author’s complete history to categorize them in social, political, linguistic, cultural, and economic dimensions.

Chapter 5 examines the machine learning techniques used and considered for this thesis to achieve abstraction. It covers options for summarization of expressed content, characterizing and prototyping users, and discovering personality traits. The algorithm Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) is discussed at length, as it was used in many of the experiments.

Chapter 6 provides a reflection on the problem of online impression formation given the experiments and their results. Chapter 7 builds on this author’s reflections by outlining the results from an outside panel of domain experts who assessed the individual works and the overall direction. Chapter 8 concludes the discussion.

1.5 Contributions

This thesis offers the following contributions:

- A description and formalization of the problem of online impression formation of strangers through the lens of data portraiture and abstraction. This can be used as a basis and working vocabulary to approach new data portraits.
- New interfaces for browsing existing individuals and crowds by demographics, opinion, network behavior, and emergent semantics. Each combines the history of the user with statistics and machine learning to achieve the abstraction effect.
- A wide variety of representations for textual data and its aggregation, including geographical metaphors, numeric scoring, naming existing social prototypes, and statistical visual language such as stacked graphs,.
- An aggregation of heterogenous information into a single data portrait from a wide variety of data sources using natural language tricks and machine learning.
- An algorithmic classification of perceived user intention in social and promotional axes using their socially meaningful network-based behavior as opposed to content analysis.
- A toolkit of techniques for abstracting social data and an outline of their usefulness and past results.
- Insights for future researchers who wish to tackle problems in online impression formation, such as issues surrounding representation and observation, distortion effects when visualizing humans, cultural and societal complexities and affordances, control of the data, and alternative paths.
- Various papers, talks, press, public exhibits at museums and research fairs worldwide, and Internet-accessible works with millions of hits.

2. BACKGROUND

This section examines how we currently understand strangers by a literature review of impression formation in face-to-face and online contexts. Prototype theory and interaction scripts are presented as core concepts for impression formation and interaction with strangers in both worlds. Differences in cost and affordances between the two worlds are discussed, setting up the goals of this thesis and painting future possibilities.

2.1. Understanding and dealing with strangers.

In order to think about the online presentation of strangers, it is first worthwhile to examine the theories from sociology about face-to-face communication and determine the connection to CMC. We review findings from three key sociologists, Simmel, Goffman, and Bourdieu, to orient the future discussion.

GEORG SIMMEL

Posting and replying to comments online is a strange method of communication considering that participants' identities exist as opaque and uncoordinated fragments of text. How can we make sense of a comment without knowing anything about the author? The same question has been asked of interactions with strangers on the street. We know nothing of the stranger, yet we can efficiently navigate conversations with shared boundaries.

Simmel hypothesized that we see another stranger as a generalized instance of ourselves. Because we cannot “*fully represent to ourselves an individuality which deviates from our own*” (Simmel, 1910), we extrapolate from our own interworking and expectations to guide iterative communication. With each speech act, we generalize, specify and typecast individuals into the categories by which we aprioristically understand the world. Simmel calls such categories “*human types*,” and it is precisely these categories that enable us to interact with each other. He postulates that society would be impossible without doing so, because we are always working within “*relations [of] varying degrees [of] incompleteness*” (Simmel, 1910).

Public online space is a dearth of cues. Is a controversial response just a “troll” trying to incite a flurry of criticism for fun, or instead a legitimate expression of an unpopular point of view (Golder, 2003)? Because CMC's level of anonymity invites trouble (Kiesler, Siegel, & McGuire, 1984), one needs to be able to quickly assess another's intentions to avoid wasted time, money, and energy. Worse yet, the fragments of given text only provide small insights into the commenter's position, limiting the judgement on which to base our potential reactions.

For example, consider the following comment from the New York Times website in the May 1st, 2009 article titled “Souter’s Exit Gives Obama First Opening”:

YAY!!! Lets get a woman on the court again!!! I hate the one woman Supreme Court. Everyone knows that women are the MOST flaming liberals, so lets see Obama do what his party wants and get LIBERAL!!!

— Woodtsunami, Cincinnati, OH

This comment is difficult to interpret. Is Woodtsunami being sarcastic using insidious language, or instead being strategic about getting more liberals to serve on the Supreme Court? Without a “*multiplicity of psychic contents*,” (Simmel, 1910) attaching form to this comment and commenter becomes a function of the relative distance from the argument’s position and choice of words to that of the interpreter. Is this comment something that I might say, in which case Woodtsunami is like me, or of a diametrically opposed category? Without an apparent social geometry to help structure the dynamics of interaction, comprehension requires filling in very large gaps using only hints provided by the larger culture.

PIERRE BOURDIEU

Simmel (1910) argues that we formulate human types to facilitate interaction under the (permanent) presence of uncertainty. Bourdieu theorizes what might guide these human types, and how their existence is affected most by the upbringing of an individual. He uses the notion of the habitus to explain how we structure the world, and how it structures us.

The *habitus* is “*the durably installed generative principle of regulated improvisations*” (Bourdieu, 1977). It can be seen as the “*unity hidden under the diversity and multiplicity of the set of practices performed in fields governed by different logics and therefore inducing different forms of realization, in accordance with the formula: [(habitus) (capital)] + field = practice*” (Bourdieu, 1979). That is to say, the habitus represents the various ways that society structures and differentiates itself through schemas, dispositions, and taste. It is a primary function by which we select the audiences we respect and whose opinions we find amenable. These principles are installed subconsciously during childhood as a direct result of the positions and practices of one’s parents. Secondly, they are altered through life as a result of education and society, amongst other systems and agents. As we are structured by others, we obtain and repurpose these structures to use on others in our world.

We rely upon our “*matrix of perceptions, appreciations, and actions*” (Bourdieu, 1976) to function predictably and make sense of the stimuli we encounter. As much as our habitus governs the furniture we buy and the music we listen to (Bourdieu, 1979), it also guides our preferences for which online forums we participate and how we choose to engage with them.

A large concern for public discourse is whether a heterogeneity of habituses leads to a strict and codified heterogeneity of participation across cultural and subcultural fragments. Recent research has illuminated discrete silos of information transmission and viewership even across a very large number of blogs, highlighting the pitfalls of homophily in garnering a revolution of collective participation (Adamic & Glance, 2005). By looking at link segregation, information types and interaction on such blogs, we can hypothesize about the shared habitus of a given set of interlinked social spaces. It might be possible to leverage the intersection of habituses as a bridge to connect otherwise disconnected nodes.

Even within a single online space, we still consolidate and process multiple habituses. When trying to mentally position and understand a comment, we look for cues to “*situate others in the hierarchies of age, wealth, power, or culture*” (Bourdieu, 1976). As stated above, many of these cues in CMC are missing. Instead of using attributes traditionally available face-to-face, we must infer their larger structural logic from the few preferences the commenter implicitly or explicitly conveys. Such preferences might be in word choice, pseudonym construction, grammar, social conduct, political viewpoint, the depth of distinction made (Bourdieu, 1979), alliances with fellow contributors, or viewpoints held (Bourdieu, 1976). Once the larger logics are inferred, we can now estimate how our practice might interact with theirs. Bourdieu’s revelation yields the content and form by which Simmel’s human types can be made more concrete. Michèle Lamont expands on Bourdieu’s ideas to show how the habitus manifests itself through symbolic boundaries we use to separate others from ourselves and our class; boundaries like socioeconomics, morality, and religion (Lamont, 1992). These boundaries underly our arguments online, but are rarely explicitly stated. It might be useful to make such distinctions more explicit in presenting a user’s identity to compensate for the small amount of information present. Working within the framework of the habitus is a vastly untapped area of CMC that could help actualize public discourse in a way that might be better than face-to-face (Holland & Stornetta, 1992).

ERVING GOFFMAN

Goffman uses a metaphor of dramaturgy to describe our interactions with others in daily life. His insights are useful in thinking about each element of impression formation and management, where performances ultimately are bounded technically, politically, structurally, and culturally (Goffman, 1959). Each performance is composed of “*all the activity of an individual which occurs during a period marked by his continuous presence before a particular set of observers and which has some influence on the observers*” (Goffman, 1959). The actors are those involved in the interaction, giving their performance on the front-stage to the observing audience. The performance yields our impression of the actors or, by their reaction, the audience. The scripts may be provided a priori by the society or culture in which the actors belong, or they may be more improvisational.

In both conditions the goals of the actors shape the given performance. One of the main ways they shape the performance is based upon the mask they wear, signifying “*the role we are striving to live up to--this mask is our truer self, the self we would like to be*” (Park, 1950). In turn, the audience interprets the performance by the fragments of identity displayed.

Goffman sees performances in two distinct regions, the front and back stages. In the front stage, we conduct ourselves according to the restrictions of the roles we are playing and larger social mores. Politeness in the front stage can be seen as the attempt to mitigate tension and to act as a social lubricant when masks collide with incompatible behavior (Watts, 2003; Brown & Levinson, 1987). In the back stage, we shed our rigid mask and expose the secrets that would disrupt the believability of the character. We don't care about politeness in the same way; social mores are eliminated or replaced by a looser set. It is less clear if the need to save face or maintain a division between regions exists on the Internet as in face-to-face performances. When interaction occurs at such a large scale that we are effectively anonymized, the consequences and perceptions of tension are reduced. Loyalty and discipline, “*attributes required of teammates if the show they put on is to be sustained*” (Goffman, 1959), are hard to coalesce when relationships exist only for a flash of time. Because the consequences are so important to interaction, Goffman's analogies help support the cause for user history to be more available in deciding with whom to engage in the first place.

There is no doubt that people present distinct masks in their interaction online. Trolls, vandalism fighters, answer persons, discussion catalysts and flamers are names of masks who reliably exhibit specific characteristics in content and method of performance and their relation to others (Gleave, Welser, Lento, & Smith, 2009). Even though most online interactions have little to no exchange (Lampe & Resnick, 2004), communities such as Wikipedia and certain Usenet newsgroups contain members who build and maintain stable identities in the context of the group (Welser, Kossinets, Smith & Cosley, 2008; Golder, 2003). It is these contexts in which the presentation of user history can be most effect. Interestingly, subcultures of interaction patterns, terminology and inside jokes tend to emerge in CMC (Sproull, Kiesler, & Zubrow, 1984).

Understanding the templates of scripts is important in presenting subcultural capital (Thorton, 1996), where deviations expose the fragility of a community member's mask and garner recourse or retribution from the community. For example, consider the following interaction in Slashdot:

Anonymous Coward, post #26576067, moderated value: Score 5, Funny
In Soviet russia, System operates YOU!

Selfbain, post #26576941, moderated value: Score -1, Troll
Look this joke is very simple. If it doesn't make sense when you reverse it, you're doing it wrong. If we reverse your joke we get: You operate system.

MindKata, post #26577141, moderated value: Score 5, Funny

"Look this joke is very simple. If it doesn't make sense when you reverse it, you're doing it wrong. If we reverse your joke we get: You operate system."

system operate You: get we joke your reverse we if wrong it doing, you're it reverse you when sense make doesn't it. If simple very is joke this Look.

I still don't get it? ;)

We can see that while the community approved of the original instance of the Russian Reversal joke, a joke popular on the blog, another member tried to rebuke it. Because the audience decided the original post was a believable performance on slashdot, they rebuked the rebuke by moderating it down and collectively assigned Selfbain the mask of a troll. Another member further admonished the troll by making fun of the rebuke in the pattern of the joke itself, only possible because MindKata had such a strong understanding of how to perform the script. This in turn commanded the likes of a standing ovation from the audience by the receiving the highest moderation score possible. We can see that the community has strong emergent fronts, and when members do not correctly perform their intended mask, the audience responds with an acute awareness. However all of this could be tinted if we had more primary access to Selfbain's history, altering a permanent "mask" that showed her usages of the joke. Such as task is not easy: humor is notoriously difficult for computers to recognize (Mihalcea & Strapparava, 2006), in part due to the strong cultural narratives latent in jokes (Ruch, 1998).

2.2. Online impression formation and identity

Prototypes, performances, and habitus are communicated and interpreted in part through the emission of cues (Kunda, 1999). These cues and social signals emerge from a diverse and rich feature space in face-to-face communication. As we use paralinguistic, linguistic, semantic, and other methods to understand and characterize people (Ekman & Keltner, 1997; Goffman, 1959; Lea & Spears, 1995; Zebrowitz, 1996), the limited bandwidth of text (the web's *lingua franca*) makes it hard to imagine is it even possible to accomplish the same tasks online with reliability and ease. However, the diminished quality of these cues online does not mean that strong impressions are not possible to make. The ability to ascertain the other online has different qualities and implications, and advancing technology makes it possible to draw conclusions automatically that would have been impossible just yesterday. Here we review some of the issues and possibilities with forming an impression textually.

SIGNAL LOSS

We are devoid of many cues in online textual communities. Fortunately, semantic and linguistic cues are still first-rate, issues of ambiguity aside (Zinman & Donath, 1999). However, its differing paralinguistic cues afford new possibilities not usually thought of -- cues that may be more legible in the all remembering virtual world than the physical. Which authors we reply to, when we do it,

our temporal rhythms, what we are rated by others, the percentage of our messages contain typos -- these are all signals that we may not think much of when participating at any one time, but taken as a whole may stand in to aid first impressions (Donath, Karahalios & Viégas, 1999).

But just as Goffman distinguished between the “*expressions given*” and the “*expressions given off*” (Goffman, 1959), it can be difficult to discern what is authentic and what is controlled or manipulating online. We allow ourselves to be visible online mostly at our own discretion. Except for identity theft, automated account creation, or other frustrating aspects of modern data living, we shed our bodily restraints behind when we employ our keyboards and mice. With each keystroke we define our alter ego, creating a persona with more gusto than may be possible in real life. Unfortunately, electronic media operate at lower resolution than we might like. This lower resolution permits certain kinds of deception to occur (Donath, 1998), and it also limits our ability to interpret the actions of others. They might be polite, which requires control, or they might be suave. In the physical world, have more types of cues to stereotype others into a form that we can process (Goffman, 1969; Bourdieu, 1979). Most cues we give away unconsciously; our clothes, gait, sociability, job, word choice, and furniture are predictive of socioeconomic, cultural, and educational capital (Bourdieu, 1979; Bonvillain, 1993). As discussed above, this is not necessarily a problem; it is a solution that makes interaction with large populations possible (Simmel, 1909; Simmel, 1950).

The relationship between impression and identity is a very entangled one, as self-presentation is simultaneously easily manipulatable and revealing of the unconscious. This only becomes exaggerated online when so many cues may be missing or easily faked, making it more difficult to feel confident in the prototypes we make online. For some, this is a long standing dream (Hiltz & Turoff, 1978). The virtual world promises for ideas to stand on their own, permitting interaction to be truly a meeting of the minds. Our minds may not match our bodies or conditions, and the ability to detach what might be irrelevant is very appealing. Yet the qualities of a textual medium do still afford inferences outside of the qualities of one’s ideas -- and the ability to make these impressions are central to a human being situated within a sociological context. They are not “*interpersonal noise*”:

“Computer-based teleconferencing is a highly cognitive medium that, in addition to providing technological advantages, promotes rationality by providing essential discipline and by filtering out affective components of communications. That is, computer-based teleconferencing acts as a filter, filtering out irrelevant and irrational interpersonal ‘noise’ and enhances the communication of highly-informed ‘pure reason’--a quest of philosophers since ancient times.” (Johansen, Vallee, and Collins, 1977)

Can filtering out interpersonal “noise” really lead to “pure reason”? It is known that CMC leads to less status effects and more equal participation (Kiesler, Siegel, & McGuire, 1984; Spears & Lea, 1994). It also tends to invoke hostility and lead to discussion of more extreme points of view

(Kollock & Smith, 1999; Siegel, Dubrovsky, Kiesler, & McGuire, 1986). Further, just because we cannot smell the remote speaker does not mean they do not transmit interpersonal characteristics. Hancock and Dunham (2001) note that in CMC “*a partner’s choice of descriptive devices (e.g., geometric vs analogic descriptions), communicative style, and paralinguistic (e.g. use of emoticons, punctuation, capitalization, etc.) all [provide] potentially impression-relevant information.*”

Users further compensate for signal loss by culturally increased expressivity in language adaption. Paralinguistic, such as the use of emoticons, is one method of increasing expressivity. But various acronyms, netiquette and other designed or emergent stabilizations in practice have moved from the computing subculture (Sproull, Kiesler, & Zubrow, 1984) to the common place (Ito, Okabe, & Matsuda, 2005). While unintentional non-verbal signals are always lost when using mediated channels, the common understanding of this loss has led to a cultural expectation to re-encode the intended signals in a form that is compatible with the active channel. With enough experience and hard lessons, the average user now deeply understands that sarcasm can be lost and that ambiguity is all too easy to transmit. They have learned that signals should be carefully constructed to facilitate interpretation and thus impression formation. These lessons culminate in a culture of preemptively injecting disambiguation into the message, most commonly using emoticons or acronyms that directly refer to emotions (e.g. LOL, ROFL, etc). In this way, text-based CMC can be seen as co-adaptive, whereby norms get transferred and are shared amongst various individual technologies (Mackay, 1990), as was predicted over thirty years ago by Hiltz and Turoff (1978):

“With time, it can be expected that users both individually and as a kind of ‘collective ‘subculture’ will develop much more skill as well as some shared norms and understandings about etiquette and level of participation, such that the observed behavior will be much more ‘regular’ or ‘predictable’ than has occurred in field trials thus far.”

Because these new avenues to consciously signal are culturally based, their predictability is relied upon in an efficient manner. The improvisational nature in cultural memes facilitates impression formation by inventing the ways to communicate that are otherwise too lacking.

ANONYMITY AND PSEUDONYMITY

Anonymity is the obvious affordance of CMC. Intimately tied to the goals of cyborg theorists like Haraway, anonymity facilitates shedding the preconditions behind one’s position in the world. Perhaps ironically, the need to find something to judge leads to an increased reliance on the few remaining social cues, such as status or role, to form an impression of the remote user (Spears & Lea, 1994; Lea & Spears, 1995). Hancock and Dunham (2001) note that “*CMC retards the rate at which impression-relevant cues are exchanged during social interaction, rather than simply reducing or eliminating the amount of such information. Communicators are assumed to take an active role in forming impressions through text-based information.*” These cues are needed to interpret the stranger. According to the Social

Identification/Deindividuation (SIDE) model, judgement is based on group similarity or difference without sufficient individuating cues (Lea & Spears, 1992). Note that the lack of anonymity is not the same as gaining a deep insight into an individual: many have very small digital footprints. Anonymity is also not the same as pseudonymity, which could provide a strong insight into a character without any way to identify them in the offline world. The definition of pseudonymity is expanded here to use a data body as an identifier.

A perceptive user might be able to manage their individuated role, and subsequently foster their designed pseudonymous impression. Yet, this can be hard to maintain over time, as one tends to leak information that either dislodges existing masks or previews glimpses of a different self (Donath, 1998). This is not to say that an online persona cannot maintain some large differences from real life. In an experiment by Walther (1997), geographically dispersed participants consistently rated the attractiveness, productivity and affection of the remote user higher in CMC than in FtF conditions. Without enough evidence, aspects of a persona get filled in with exaggerated attributions. Furthermore, *“impressions can become more intensified over time as participants engage in selective self-presentation and cognitive reallocation and as intensification processes such as behavior confirmation begin to operate”* (Hancock & Dunham, 2001).

A major affordance of anonymity and pseudonymity is multiplicity; one explores and projects several selves within different communities. Turkle (1999) notes *“it is not unusual for someone to be BroncoBill in one online community, ArmaniBoy in another, and MrSensitive in a third.”* Each of these personae are real, to a certain extent, in that they are aspects or fragments of the mind of the user regardless of ephemerality or lack of an identifier. This not unlike the real world, where we choose our mask according to the social situation and the self we wish others to believe (Goffman, 1959). However, unlike the real world, our virtual mask is often consequence free. It allows us to try out behavior modifications to understand the reaction, or to actualize elements we hide in real life for fear or impossibility. In this sense, anonymity, persistent pseudonymity, and multiplicity afford cyber-psychotherapy. Turkle (1999) offers the following insight:

“People who cultivate an awareness of what stands behind their screen personae are the ones most likely to succeed in using virtual experience for personal and social transformation. And the people who make the most of their lives on the screen are those who are able to approach it in a spirit of self-reflection. What does my behavior in cyberspace tell me about what I want, who I am, what I may not be getting in the rest of my life?”

What better place to transcend physical identity than in a *“consensual hallucination”* (Gibson, 1987)? Much of the hope for manipulable appearances comes from the adage of virtual reality, where some days you might choose to be *“tall and beautiful; [and] on another you might wish to be short and plain. It would be instructive to see how changed physical attributes altered your interactions with other people”* (Krueger, 1991). We know that appearance, virtual or physical, does indeed change behavior (Donath, 2001; Bailenson et al, 2001). This is perhaps why so much of identity fantasy

online involves gender play (Bruckman, 1993; Turkle, 1995), because it is the most understandable and well defined set of roles which we know, and have a difficult time decoupling from our native gender. When we switch sexes, and the remote participant has no idea, we enjoy a gendered response implausible in real life without considerable surgery. It is also much clearer what to expect, and thus want, from switching genders than from abstract characters like “*Photon the Clown with a 95-foot-long triple penis made of marshmallows*” (Sheff, 1990).

The multiplicity itself is not a downfall for impression formation. As long as the characters people play are predictable and understandable, society can function (Goffman, 1959; Simmel, 1909). After all, we all play different roles according to the context -- there is no one self (Goffman, 1959). Thus each persona can be effective in their goals and understandable as a complete whole, should the actor perform a consistent character.

THE NETWORK

People online can be understood by more than just their own behavior. They can be understood in part through observing their relationships with others (Donath, 2008). Currently most online media do little to make one’s relationships easily legible to others. Facebook displays how many friends one has and the latest things they might have said to them, but the overall strength of those relationships are hidden (to their algorithms). Myspace has the concept of Top 8 friends, much to the chagrin of many thus tortured high schoolers (boyd, 2006). But no “app” exists yet that tells the world if you curse in front of your mother, or better yet, if she curses. Many would feel such information is too private for public display, preferring to reveal those aspects only to intimates. However, constraining and hiding the more meaningful relationship moments online limits the possibilities to form a better impression. Empirical evidence of a given type of behavior can be made available digitally to anyone. It is projected that a given dating profile would be assessed very differently should it reveal an abnormally complex relationship with one’s parents or past lovers. This is a unique digital advantage, although the persistence of data is not without negative consequence (Bell, 2011).

Determining emotional behavior of participants in an online community is a useful path to expose potential consequences from engagement. For example, this author recently joined a mailing list for a local motorcycle enthusiast club. One of the first messages to arrive was as follows:

STOP USING THE RIDER MC GROUP EMAIL FOR YOUR TALKING TO EACK
OUTER.
I GET ENOUGH OF EMAIL AND DON'T WANT SHIT LIKE THIS IN MY BOX!

THINK ABOUT WANT YOUR DOING BEFOR YOU SEND OR REPLY...

YES I SENT THIS TO THE GROUP SO YOU ALL CAN SEE WHAT I SAID TO THOSE
3.

Al

While this purposely public display of negativity gave concern, it was alleviated by the outpour of support for the others involved while reprimanding Al. Here my concern was not about Al, but rather how the collective reacted to his provocation. Being able to assess the normalcy of this kind of behavior directly affects the desire and feelings of safety of joining such a group. Technology should permit a more accessible way to scour already the archived mailing list to answer such a question.

SOME ISSUES WITH TEXTUAL COMMUNICATIONS

The primary medium of interaction online is text, which presents a set of challenges and opportunities when forming impressions of an individual online. It was argued above that ordinarily nonverbal cues will shift to new media channels because of a common understanding of channel capacities. This logic resonates with the Social Information Processing (SIP) model, which theorizes that while online relationships take longer to establish, “*CMC can supersede levels of affection and emotion of FtF interaction*” (Walther, 1992). This has been recently verified by Hancock et al., who provide empirical support to show that indeed mediated textual communication can carry a large percentage of what normally would be nonverbal emotion and content (Hancock, Landrigan & Silver, 2007).

So not only can we understand the rich emotional expressivity in text messaging, a variety of other impressions can be formed. The writing capability of any one individual proxies Bourdieu's habitus in word choice, background education, cultural differences, and choice of subject matter. As argued above, any one person's communications are subject to the same structuring principals in a society online as offline. We know what our education and life experiences have taught us, which comes across in our opinions and biases. Similarly, we recognize and seek out those with a similar background due to homophily (Adamic & Glance, 2005). Textual communication affords this inference of background, aiding our ability to socially navigate the web.

Thus it only makes sense that certain websites become communities or ghettos for likeminded people (boyd, forthcoming). YouTube early on attracted a particular kind of demographic, which in turn spurred more content to reflect that perspective and habitus, which in turn reinforces the same community presence. While YouTube may use video as its primary communication, the effects of homophily transcend medium. Just as subcultures find their own signals and justifications for their identity (Hebdige, 1979; Thornton, 1995), they adopt their expression to

the new medium in a similar spirit to nonverbal communication in Social Information Processing theory. Just as the symbols of the English language afforded emoticons, the effects manipulating capitalization and spelling were apparent to hackers early on, these trends have shifted and now signal a different set of demographics such as the AzN community (Hudson, 1996; urbandictionary, 2011). The natural tendency to want to express identity means human creativity can ultimately triumph over mediation channels. With it comes an increased ability to make an accurate impression so long as the onlooker is aware of the communication styles of different communities.

∞. Section summary

This section has reviewed concepts from other disciplines in how we understand and interact with strangers both on and offline. The differences in affordances online lead to new qualities of trust, identity, consequence, and the ability to gauge behavior patterns. We have seen that many of the cues that are absent in real life are made up for with creative usages of new media, and that many of the cues found in the semantics or sociolinguistics of speech are well translated textually.

3. RELATED WORK IN COMPUTER MEDIATED COMMUNICATIONS

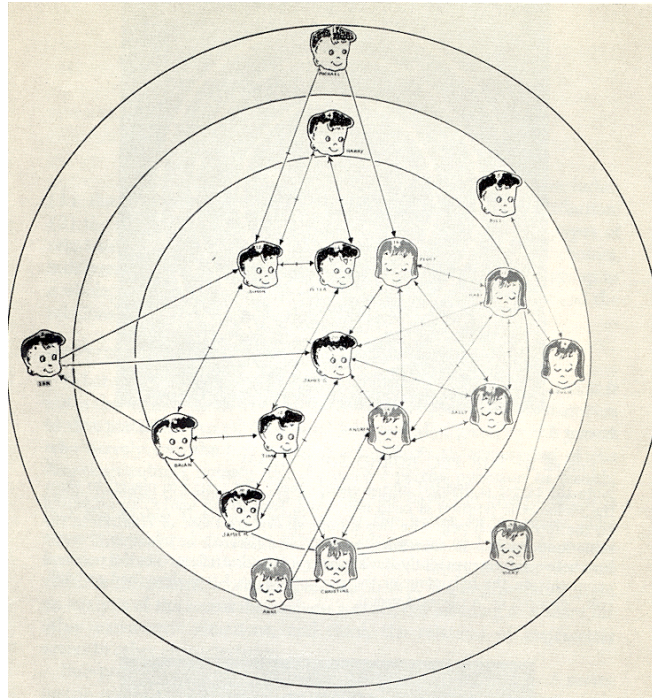
Researchers in CMC have build a variety of tools to mine and visualize the artifacts of online presence. This section reviews the relevant literature for projects and concepts that question how we may understand social spaces via the characters therein, differing methods of slicing individual's histories, and other related questions in the visualization of social data.

3.1. Making sense of social spaces

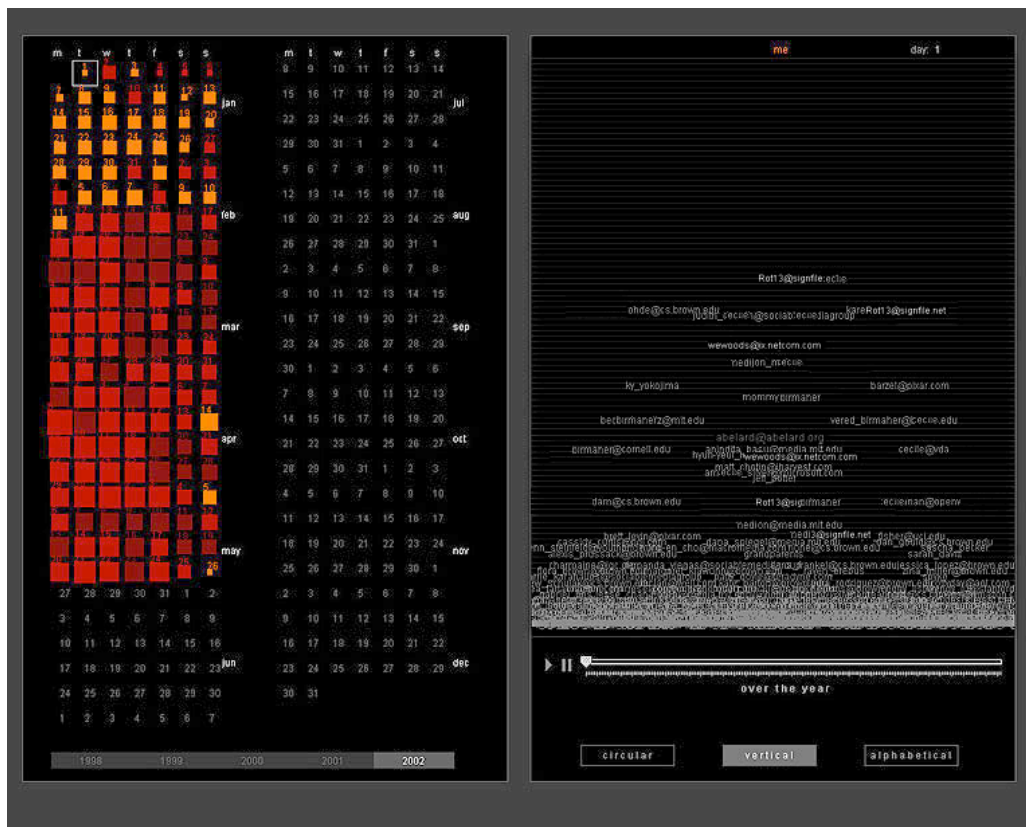
Interaction in persistent media can be analyzed structurally or semantically. The semantics relates to what is being said in the communication instances, and the structural analysis relates to the metadata of an instance. Often semantics can be troublesome to interpret accurately due to difficulties in natural language processing, and is thus recommended to approach with extreme caution. Conversely, structural-level features are far more clear-cut. Thus it should not be surprising that structural features have received much attention in social media design; they are able to characterize ongoing exchanges by proxying sociometry, or the social distance between individuals. Principal structural features include the frequency of interaction, the edges in a communicative instances, and when each interaction occurs. These structural features are separate from the structural context defined in a future section. Instead, this section of the chapter examines the potential for structural features to enhance media by providing socially relevant contextual cues in a communicative space.

Most CMC media specify the authorship and timestamp of each message. Occasionally they will additionally expose an implicit (e.g. Twitter) or explicit (e.g. Google Wave) reply structure, which demonstrates a relationship or conversation between individuals³. While these are basic and necessary steps, they do not reveal overall patterns of behavior or how the players evolved over time. But they should; CMC affords untangling complex webs of relationships and temporal trends (Donath et al., 1999), and there have been some useful inroads made towards abstracting and visualizing past activity (Donath et al., 1999; Sack, 2001; Smith & Fiore, 2001; Wattenberg & Millen, 2003; Viégas & Smith, 2004; Lam & Donath, 2005; Viégas, 2006). Classical sociology is the original source of practices that map human interaction, and it employs sociograms to reveal trends in relationships, as shown in Figure 3.1a. More recently, designers have created new visualization techniques tailored for online discussion. Based on their higher-level goals, designers choose a subset of basic statistics to explore where the intended meaning and power comes from their unique combination, as shown in Figures 3.1b and 3.1c.

³ Implicit structures use social convention to signify replies to audiences, such as when one comment on a forum directly uses the name of another commenter. Explicit structures are built into the interface so that authors may indicate to which message they reply, aiding computational assessment and end-user display.



(a)



(b)

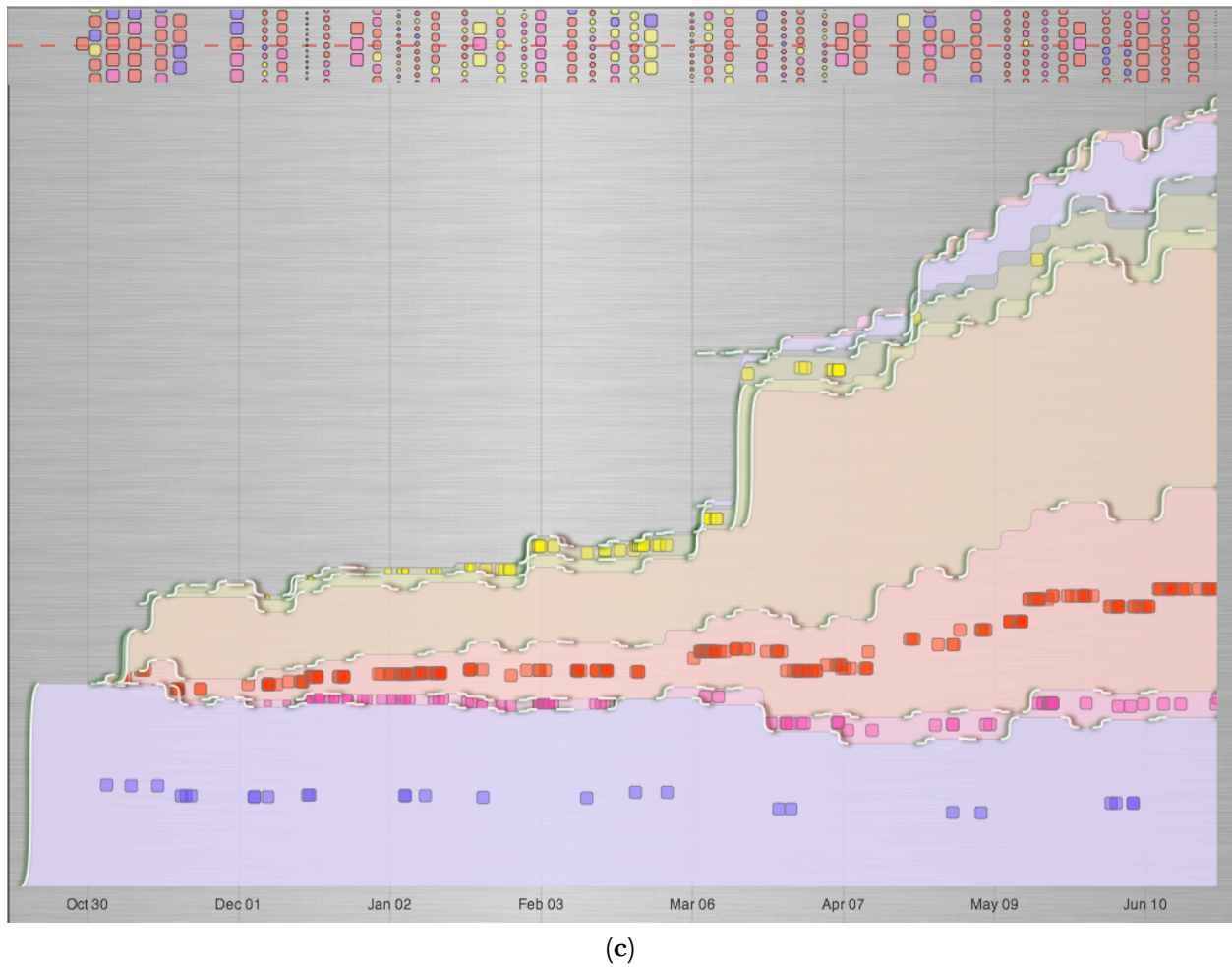
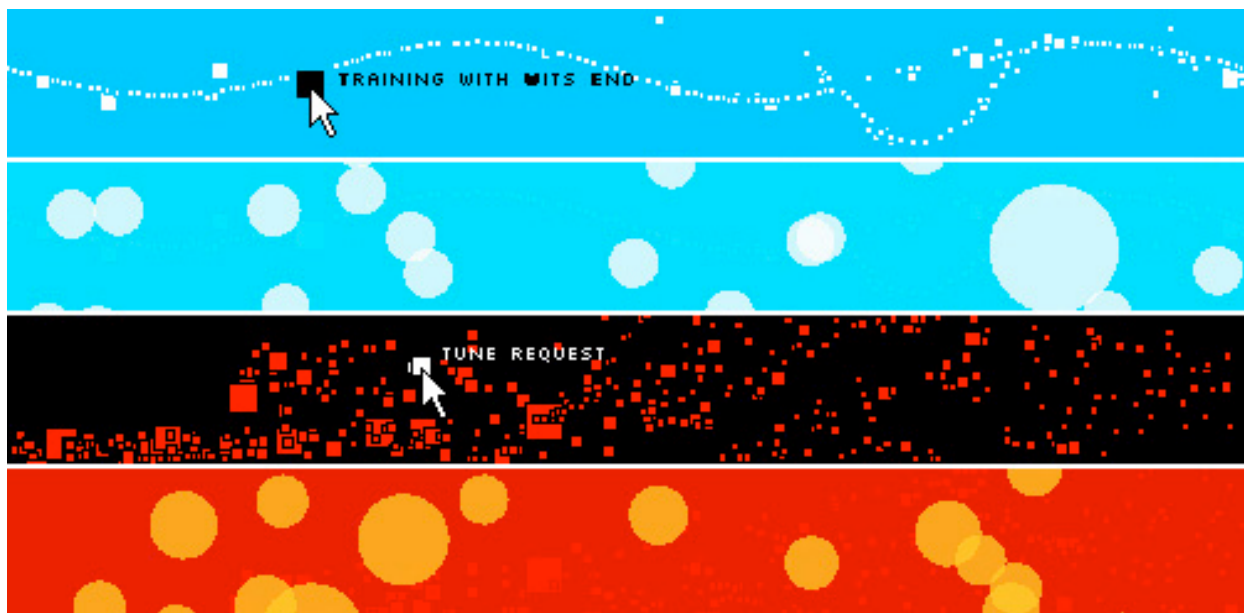


Figure 3.1. Three different visualizations of social interaction. **(a)** Northway’s (1952) “target” representation of a 1st grade classroom using a sociogram, **(b)** PostHistory (Viégas et al., 2004) visualizes contact frequency, rank, and temporal changes, **(c)** Open Sources (Zinman, 2004) shows frequency of communication in contrast to code ownership across time in a centralized software repository.

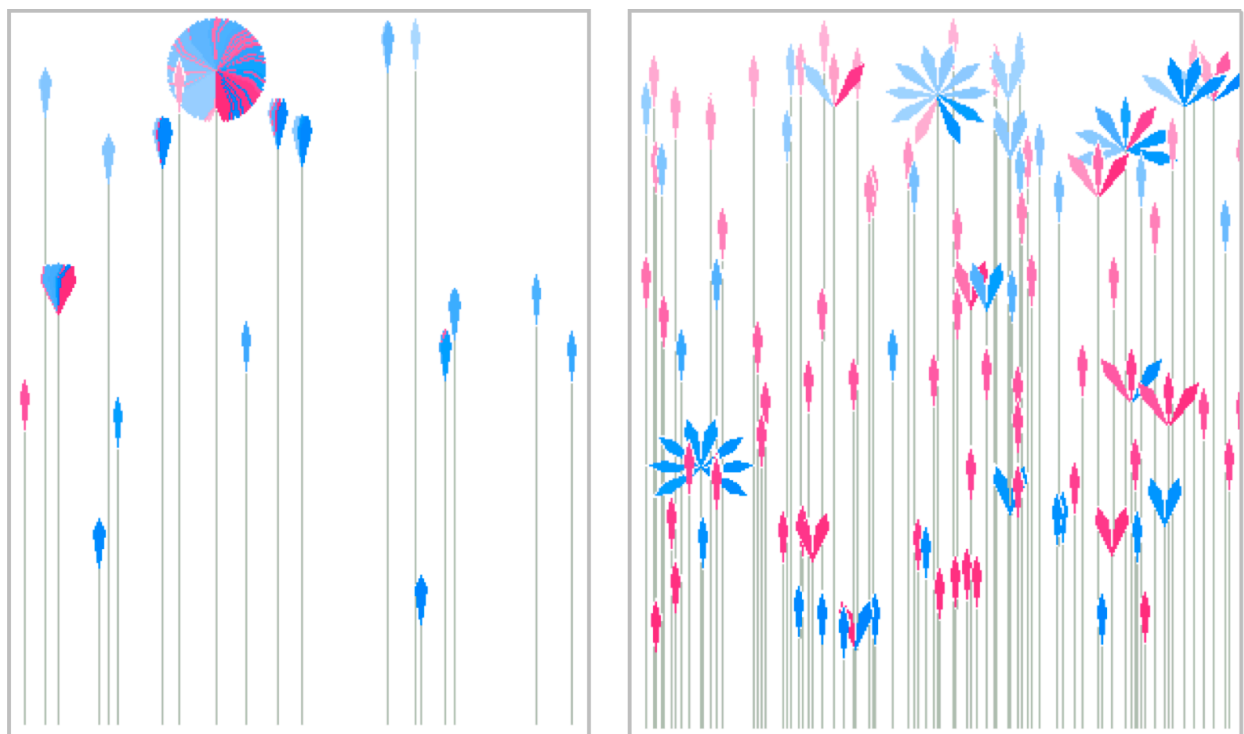
Most past work in visualizing online conversations has focused on a few high-level user goals, which, broadly speaking, fall into three categories: 1) monitoring community health, 2) discovering the main players, and 3) discovering relationships. Representative work for these goals is discussed below.

MONITORING COMMUNITY HEALTH

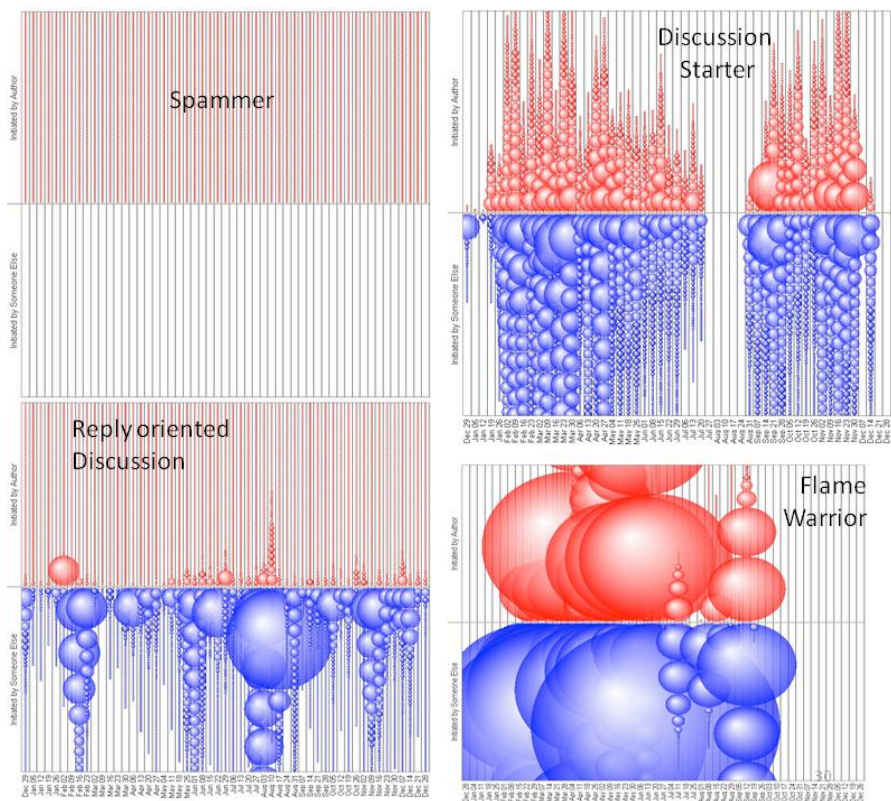
Online, like offline, communities appear and disappear with high frequency. However, CMC permits more temporary groupings due to its looser connections over many people. Thus, there are numerous scenarios where one might wish to probe the health of a community. Welser et al. (2007) used sociograms and Viégas and Smith’s (2004) authorlines to examine sociological roles in Usenet groups. In particular, they were interested in technical forums where the ratio of



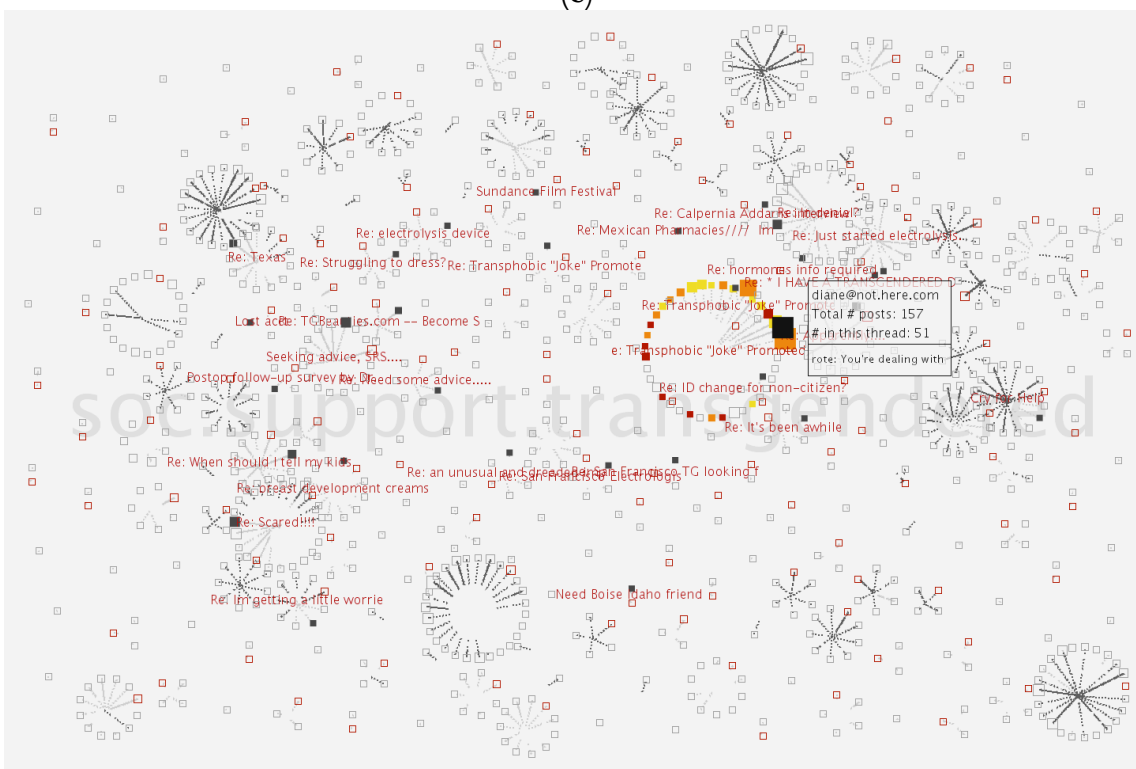
(a)



(b)



(c)



(d)

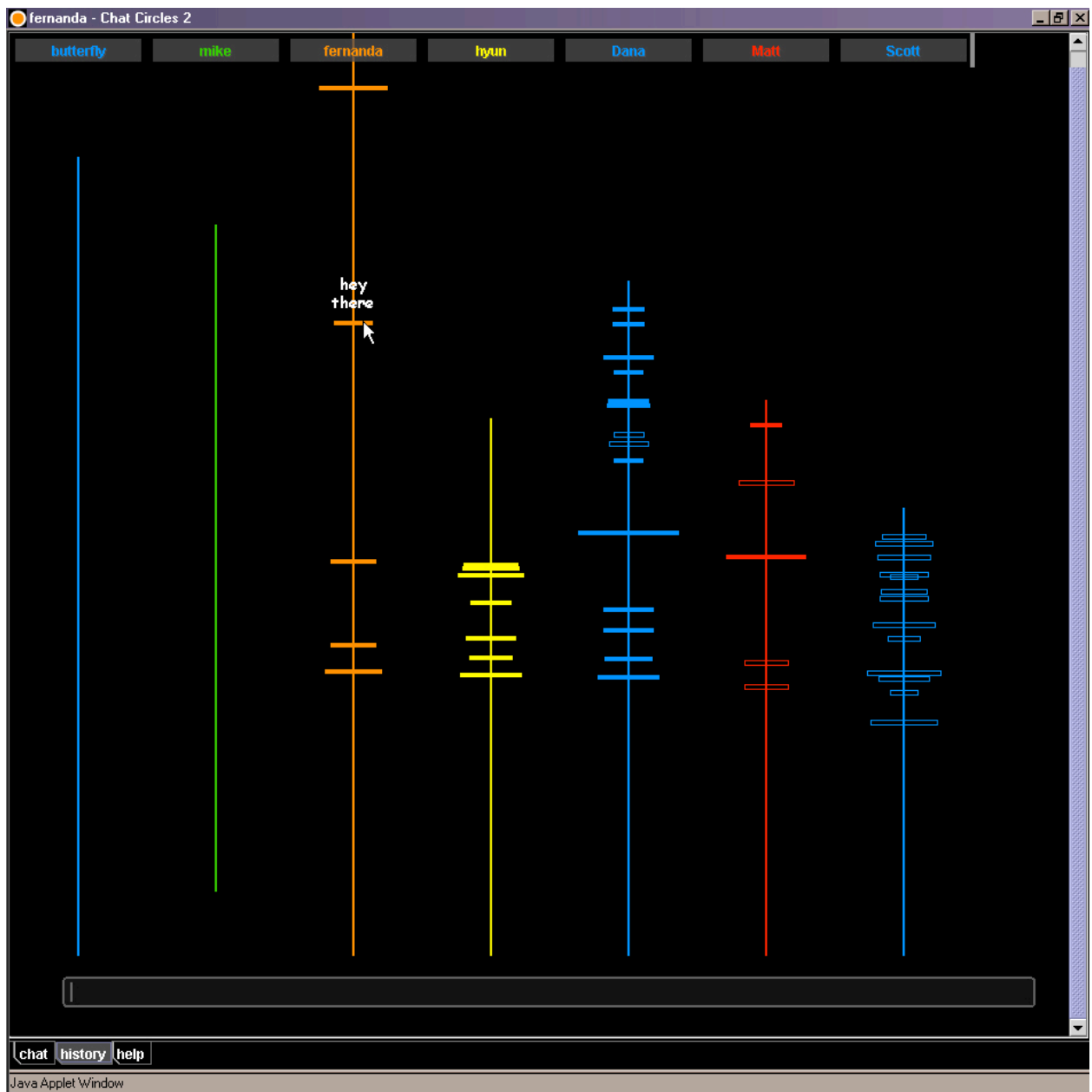
Figure 3.2 A sampling of existing visualizations for diagnosing the activity or health of a community. (a) Lam and Donath (2005) use the seascape and volcano kinetic metaphors to explore Usenet newsgroups, (b) Xiong and Donath (1999) use a garden metaphor for interpreting overall health, (c) Welser et. al (2007) use the authorlines visualization to grok the likely roles of its inhabitants, (d) boyd et al. (2002) employ semantic techniques to visualize online conversations in Loom.

“answer” to “discussion” persons have an empirical effect on discussion health (Golder 2003; Viegas & Smith 2004; Turner et al., 2005). They found that these visualizations produce distinctly different outcomes based upon the interaction styles of individual contributors, and when compared to content analysis of their actions, their roles become clear. Other researchers have been less concerned with identifying participant roles, choosing to focus on compact visuals that gestalt interactions into overall activity level. PeopleGarden (Xiong & Donath, 1999) uses a metaphor of a flower garden to demonstrate the health of Usenet discussion groups. Each user is represented by an individual flower, the color and number of petals mapped to posts and their attributes. As shown in Figure 3.2b, groups with few dominating players stand out from groups with more uniform participation. Loom (boyd et. al, 2002) achieves similar aims without the use of a garden metaphor, using more abstract representations to avoid unintended semantic characterizations of the group. Loom emphasizes the conversational aspect of a group in contrast to PeopleGarden’s user focus; it characterizes the depths of discussion trees. Seascape and Volcano (Lam & Donath, 2005) similarly focus on the conversation using stacked kinetic graphs of Usenet groups. The gestalt effect from the animation allows quick comparisons of group activity.

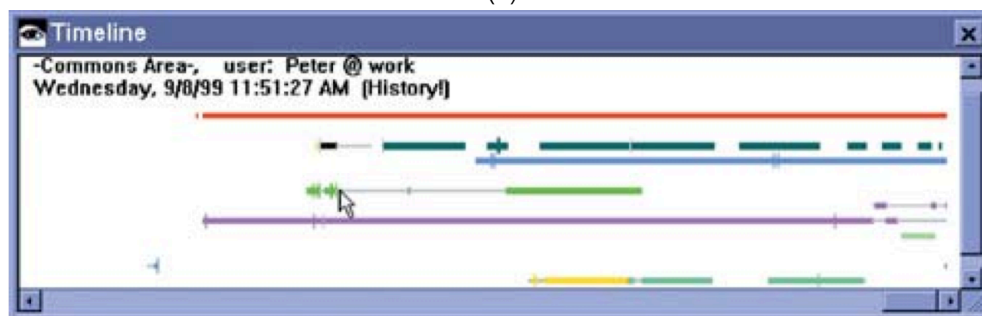
DISCOVERING THE MAIN PLAYERS

In the offline world, newcomers to a community often receive some kind of orientation to the participants and their roles, perhaps from the, director, manager, or community organizer. But public online communities are often asynchronous, allowing voyeuristic and uninvited behavior. Without orientation from a central member, it can be very difficult to know the personalities and assigned responsibilities within a space. Broadcasting a request for such information can feel awkward, making it especially difficult to enter more formal, task-oriented communities, such as open source development teams. The situation is worse in unarchived semi-synchronous channels like IRC, where the absence of history makes it impossible to grok the active discussion or participants upon entering.


CMC has the unique ability to automatically guide new participants through a social space using, among others, visualization techniques. Much like determining the health of a community, real-time top-down views can be integrated directly into media. Such views not only help newcomers, but also provide awareness of dominating voices and weak activity. Authoritative presentations of activity are known to modify behavior (Donath et al., 2000; DiMicco et al., 2004).



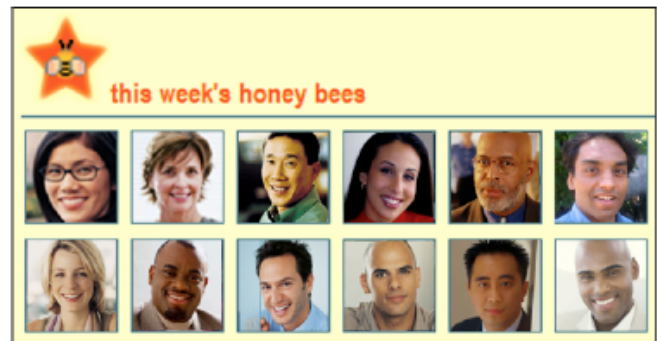
(a)



(b)

Top scorers	Top killers
 	 
1. MYAYOUSUCKGETONLINE 24000	1. DeJa Nay 104
2. CONCORD Police Captain 20353	2. Satanaska 98
3. Minmatar Control Tower 18172	3. El'Tar 93
4. CONCORD Police Commander 14358	4. Lithla Tsanov 91
5. Amarr Sentry Gun 12443	5. Soulz69 89
6. El'Tar 12193	6. Mr HappyBunny 85
7. Gallente Sentry Gun 9364	7. Darking 85
8. Kimura Masahiko 8786	8. Lee Keldar 85
9. luckymans 8663	9. MoLeH 85
10. Ragna Valdr 8207	10. VOOOBLA 82
(points in week 38)	(kills in week 38)

(c)



(d)

Figure 3.3 Visualizations that focus on users in CMC. (a) Chat Circles (Donath et al., 1999) and (b) Babble (Erickson, 2002) show timeline views of semi-synchronous participation, (c) one of many public “kill boards” for Eve Online, (d) IBM Research’s BeeHive motivates contributions by showing the most active users (Farzan et al., 2008).

Typically such work aims to isolate individual users, and show their activity streams over relevant periods of time. What activity is measured, and how it is mapped onto the visualization, is dependent on the usage style of a medium. Many past works focus on instantaneous periods of time rather than the accumulation of activity. Both Chat Circles (Donath et al., 1999) and Babble (Erickson, 2002), as shown in Figure 3.3a and 2.3b, use line-based graphs to show contribution over time in a synchronous chat medium.

Babble has additional configurations to reflect immediate activity levels to function as a social proxy (Erickson, 2002). Social proxies are intended to amplify presence as a way of increasing visible social cues online. They can be functionally constrained, such that the visualization creates pressure to return to a default social order or configuration, or communicate an ideal scenario (see Figure 3.3c). Conversation votes (Bergstrom & Karahalios, 2007), Kim (2009), and DiMicco’s (2004) work on augmented physical group meetings employ similar tactics to guide synchronous meetings using digital mirrors.

Some distributed social environments, such as those focused on productive work rather than social relaxation, require tools that focus on orienting newcomers to the long-term aggregated

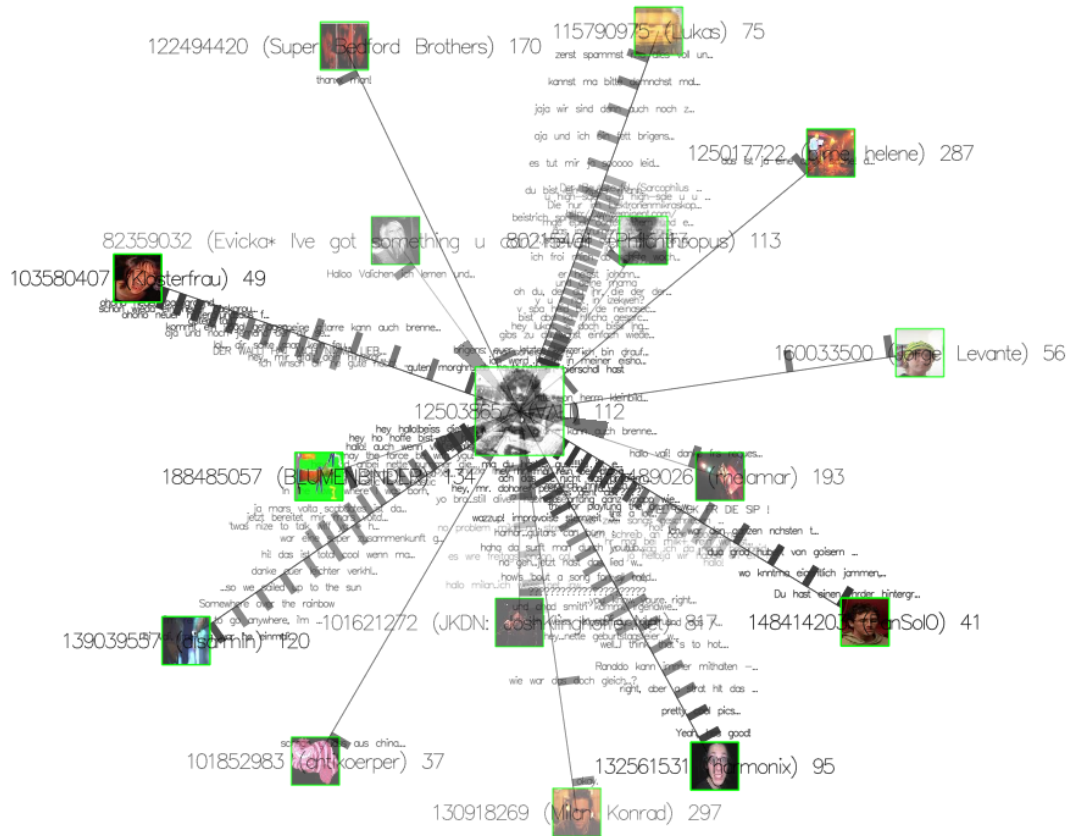
efforts. Open Sources (Zinman, 2004) peppered with short-term communication instances, visually emphasizes code contribution over time for open source communities. It takes advantage of the production of tangible value to calculate contribution as a proxy for social role. Such activity is not restricted to source code production. Many communities solicit participation through competition, prominently listing the top achievers. Traditionally existing only in video games, like community-driven Kill Boards for Eve Online (Figure 3.3d), the ranking of members is becoming commonplace in more social environments, including Foursquare's notion of mayors, Yahoo! Answers, or Busy Bees and Honey Bees in IBM Research's Beehive social network (Farzan et al., 2008).

DISCOVERING RELATIONSHIPS

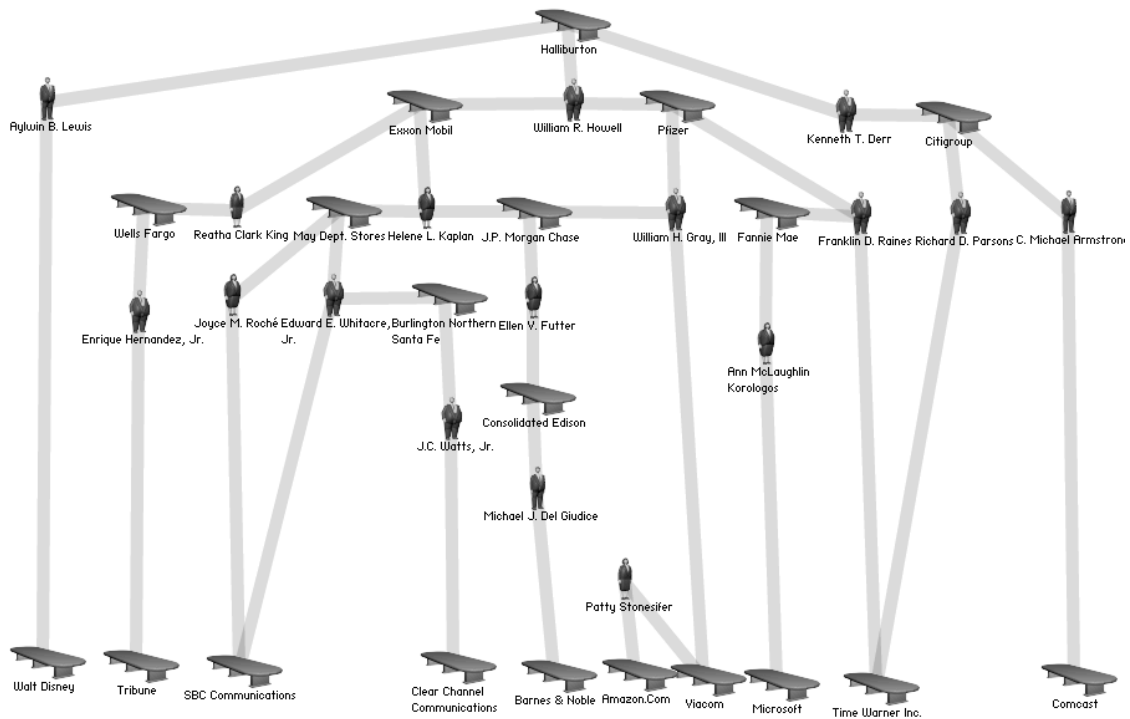
Complex social relationships are a fundamental part of being human, and wanting to understand these relationships comes with the territory. There are many ways CMC can be used to examine relationships in a community. Exploration may be for self-reflective purposes (Viégas, 2005), to understand politics and power structures (Adamic & Glance, 2005; Sack, 2007; Soroker et al., 2008), for personal information management (Nardi et al., 2002), or to simply map out the topics participants discuss (Donath, 2006).

When aggregating and visualizing relationships, often researchers first turn to social network analysis (SNA) and sociograms (Moreno, 1934; Gleave et al., 2009). SNA is a reasonable method, as it attempts to formalize relationships based upon characteristic network activity. If all communication is performed within the view of the analyst, then we can create useful authoritative graphs to demonstrate the relationships (Offenhuber & Donath, 2008). However, one must be careful to note that often much of the communication does not only occur within a single network, and any explicitly drawn social network can incorrectly represent the true relationship. This seemingly obvious fact was reinforced by Gilbert and Karahalios (2009), where a predictor of tie strength that generally worked very well using Facebook data would systematically fail for friends of friends (asymmetrically) seeking engagement, intimate relationships that use more formal media, and ex-lovers. Such problems can be avoided if humans are used in the loop to annotate or group relationships into emergent semantic categories. ContactMap (Nardi et al., 2002) serves as a central point for an integrated suite of CMC systems, whereby the user can freely associate and group together members of their social network. Its two-dimensional layout allows more expressivity than binary group membership permits.

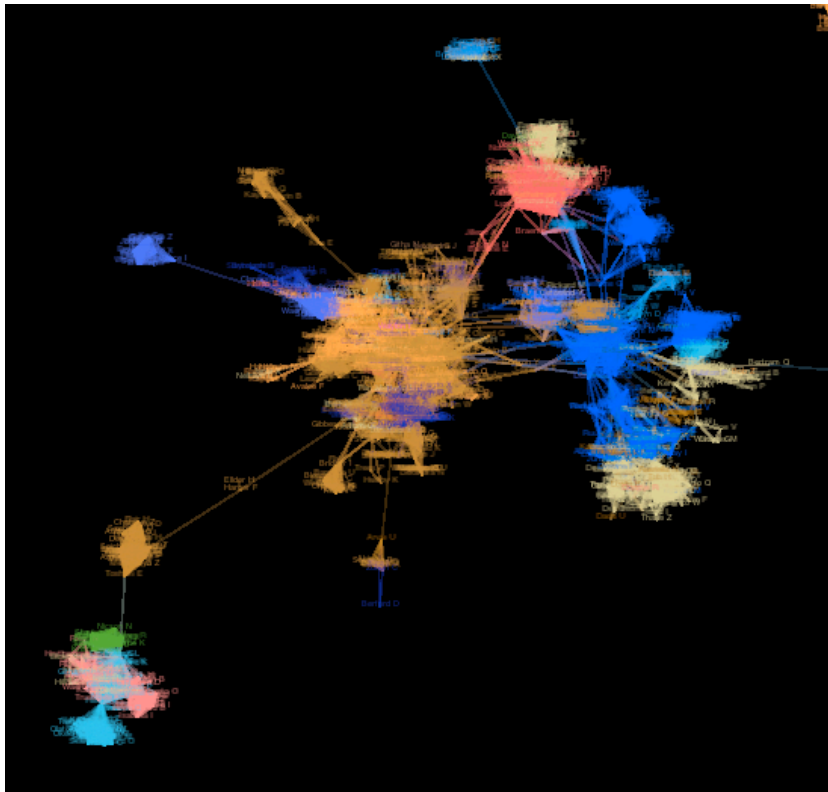
When a medium is not situated within an explicit social graph, one can be implicitly derived. This is the direction that many works have taken in the past to deal with diverse relationships and media, such as email (Viégas et al., 2004), corporate filings (On, 2004), and code revisions (Zinman, 2004). Every relationship can be characterized across a large number of dimensions,



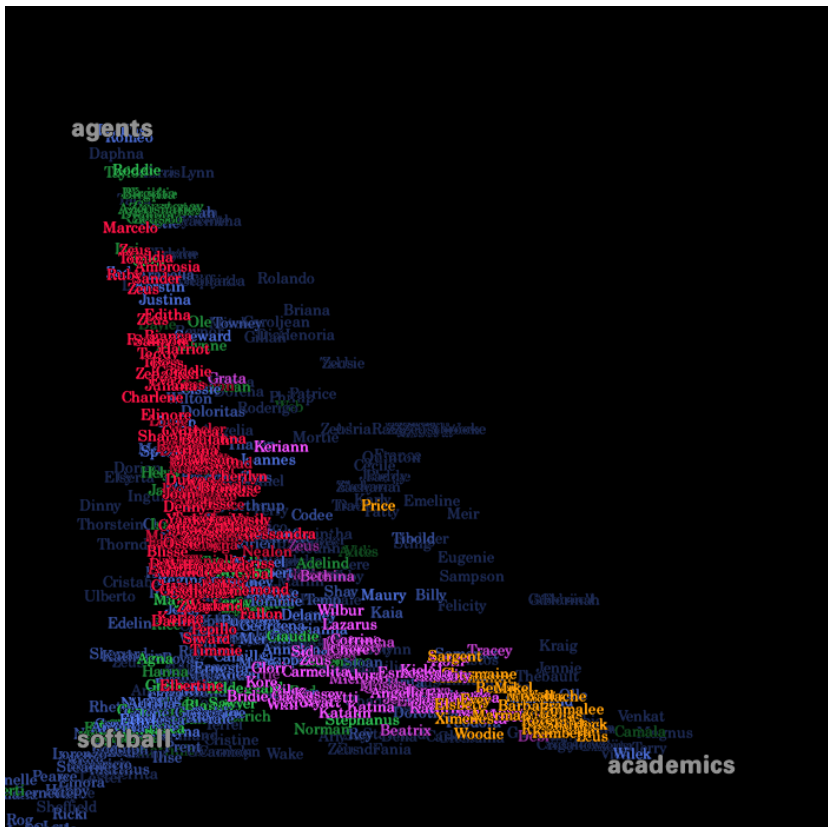
(a)



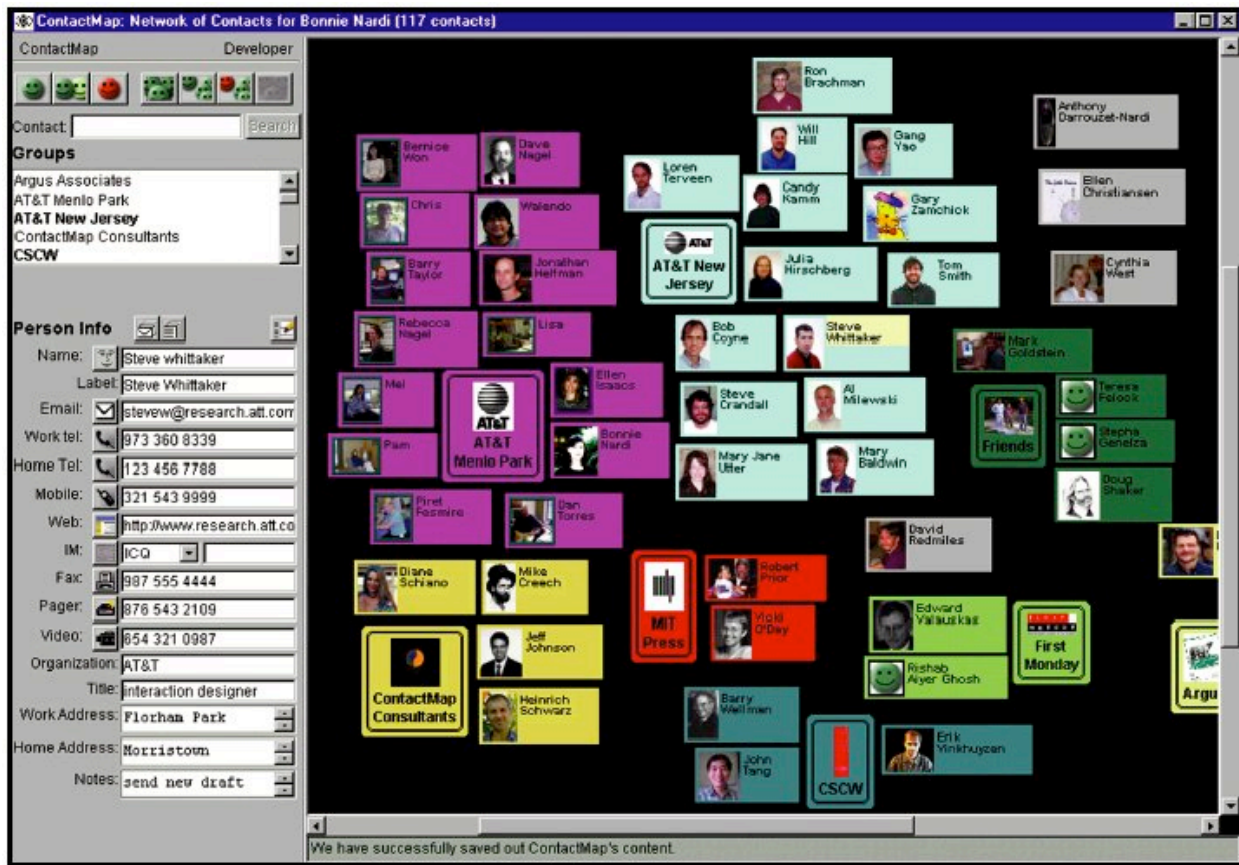
(b)



(c)



(d)



(e)

Figure 3.4 Visualizations can map explicit and implicit relationships across media. (a) CommentFlow annotates a social graph with communication instances (Offenhuber & Donath, 2008), (b) They Rule explores maps of unannotated corporate relationships (On, 2004), (c) Social Network Fragments generates implicit relationship graphs from email (Viégas et al., 2004), (d) Visual Who dynamically reveals complex temporal-relationship structures (Donath, 1995), and (e) ContactMap puts mapping and semantic categorizing into the hands of the human operator (Nardi et al., 2002).

but designers must choose a salient subset so they do not overwhelm their users. Without an explicit set of relationships as chosen by the participants (e.g. social networks), this task is very difficult. Some media afford analytics better than others: for instance, open source communities can be visualized coherently through the main outputs by the community, code and communication (Zinman, 2004), as shown in Figure 3.4c. In Social Network Fragments, Viégas et al. (2004) use easily interpretable social signals provided by the structural usage of email, as shown in Figure 3.4d. Direct symmetric and repeated email between two individuals is weighted far more highly than correspondence with multiple recipients using carbon copies. Adamic and Glance (2005) used highly reliable signals from hyperlinks to information sources as votes of confidence to reveal clear divisions and relationships between political blogs, a technique that is

more easily interpretable than On's (2004) depictions of relationships based upon un-annotated heterogeneous links of corporate affiliations.

Another approach for multi-dimensional relationships is to simply allow individuals to explore the datasets dynamically. Visual Who (Donath, 1995), as shown in Figure 3.4f, lets users place a subset of possible mailing list subscriptions (affinities) onto a two-dimensional plane, which results in a spring-based weighted layout of names that are simultaneously online and subscribed to at least one of the placed lists. While Visual Who does not attempt to characterize how much a member contributes to a mailing list, its dynamic nature allows the user *“to create many different views of the community structure and to observe the temporal patterns created by the members' activity,”* resulting in a *“multi-faceted overview of a complex society”* (Donath, 1995).

3.2. Diving deeper into the semantics of interaction

More recent work has started examining the semantics of interaction. Instead of purely looking at structural usage patterns, the communications themselves are becoming a point of reflection by more 'contextual' slicing.

SOCIOLOGICAL SEMANTICS

Sociology is fundamentally concerned with the ways in which society's fabric is constructed and used. It underpins our interaction with others, guiding sense making and community building. Society's rules also underpin most dialectics, and is the way large groups of people know how to intermingle. Here, sociological context involves exposing these rules and their instantiations. It is most likely integrated into CMC as the cues that allow us to better prototype the other using salient features to daily life (Liu, 2007).

As designers of CMC we get most of these social rules for free as humans work around mediation to perceive and project relevant cues (Lea & Spears, 1992). However, designers can create better media by understanding the limits of what gets translated or lost in CMC. Cultural issues, values, and norms impact the adaption of new technologies (Ito et al, 2005). With the proper set of affordances and sensitivities, CMC might do better than face-to-face in letting us understand the other and what the other might understand about us (Holland & Stornetta, 1992).

In real life, individual comments or ideas are never taken without some level of additional context. A large number of social cues is processed, from the environment in which the interaction takes places to the hairstyle of the interlocutor. Sociologists such as Simmel (1910) believe this is a fundamentally necessary process for society to function, as stereotyping fills in knowledge in order to make short-term communication efficient and possible. Online, we only see the small sets of words presented in each message. Smaller, long-established communities

have the benefit of repeated engagement, which maximizes communicative efficiency as the listener associates each message with those before it to provide context and reduce ambiguity. This not true of large-scale public performances which are of most concern to this thesis. The more cues of identity we can provide along side each message, the better we can color its meaning with (hopefully) helpful context.

Online identity can be a mixture of one's history online with artifacts of one's offline habitus. To the extent that we can consider and summarize past opinions, we can help conjure an image of a given habitus' structure. This possibility is further reinforced by data mining the leaky social aspects that Bourdieu (1979) has shown to be characteristic of a given upbringing, such as economic class, geographical region, and most other markers of social grouping. All such signals, along with summarization of past interactions, can be condensed into a miniaturized data portrait to be viewed alongside each communication for context and further exploration of the individual. This is an important and necessary step for CMC, where the correct portraiture and interaction suggest a kind of continuity even when audience-speaker relationships are at best discontinuous. This thesis hypothesizes that such a window into more complete interactions can facilitate discussion and connections that move beyond a limited discussion space.

Habitus also groups similarly related individuals into a collection by which their likely shared ideas or backgrounds can foster less aggressive behaviors. If the interface supports such grouping into sub-spaces for like-minded audience, it is hypothesized that 1) message quality will increase as retorts are aimed at less dissimilar positions, and 2) the number of messages will increase as more friendly and familiar audiences increase the appeal of joining a conversation.

While it might seem difficult to computationally infer one's habitus, there have been alluring inroads in machine learning to suggest it can be, at least partially, done. As Bourdieu (1979) notes that since one's habitus controls everyday decisions, aesthetics are largely telling of one's upbringing. While it is difficult to tell what furniture or clothing one owns online, social network profiles do reveal many useful preferences. Liu (2007) has shown that a kind of online identity can be calculated from such "*taste performances*" using Principal Component Analysis in such a way that the relationships between tastes are predictable and adequately stable. Furthermore, Liu has found that factors such as education and religion are able to predict much of the variance in tastes. Knowledge of such factors allow sites like Hunch to combine weakly correlated questions to provide better answers to seemingly unrelated questions. While more research needs to be performed, it is suggestive that signals exist that can be used to infer societally relevant characteristics of individuals based upon their available online data. For instance, I have used unsupervised topic modeling (Blei, Ng, & Jordan, 2003) to well separate a large corpus of text in MySpace profiles into sociolinguistically meaningful clusters (Bonvillian, 1993), as shown in Table 4.1.

Such clusters can help prototype users based upon conventional signals that humans also use to judge others in dialogue. Similarly, linguistic markers can also be used to detect controversial comments and threads (Mishne & Glance, 2006) or to characterize site usage patterns (Zinman & Donath, 2007). All such features of controversiality, human groupings of similarity, or even external features such as geography or race can be used to condition topic models using Supervised Topic Models (Blei & McAuliffe, 2007) or DMR Models (Mimno & McCallum, 2008). This allow us to predict these features by conditioning on incoming text. We can use this result to help guess the habitus, for other forms of intelligent grouping, and as input features for generative data portraiture. Cross-domain correlations stand to be the most promising, as predictors that look at a single facet of personality and preference do not play into the larger social fabric.

Topic A	Topic B	Topic C	Topic D n-grams
hey whats time long talk havent hows good talked haven goin forever ttyp	ha cute yeah today song cool thing thought didn fun made mom guess	friends real fat back life fake shit friend homies drink send parents call	eminem presents candy couture louis vuitton denim dior saddle chanel cambron chloe paddington newest styles candy couture carries balenciaga le dix motorcycle fendi spy gucci hobo

Table 3.1 Partial results from applying Latent Dirichlet Allocation to MySpace profile text.

It is important to note that humans should play an active role in shaping their presented online habitus, as the ways in which people infer habitus or group together compatible or similar people will always be more sophisticated than a machine could guess. For this reason humans should also be able to exert control when partitioning messages and annotating identities. ContactMap (Nardi et al., 2002) is such one approach. Ultimately, encouraging the involvement of the end users has the added benefit of creating a sense of ownership and pride that arises when users invest in the community.

Computed habitus and user-generated groupings are an important source of information in generating online portraiture. They help communicate persona through semantic relevancy, and are especially useful when combined with structural context as discussed above. Visualization can provide general insight into the habits of users in a way that is otherwise hidden in the streams of data.

SEMANTICS AND TIME

No messages are created in a vacuum, even if YouTube comments seem to come from another planet. Krauss and Fussell (1991) note that “*much social behavior is predicated upon assumptions an actor makes about the knowledge, beliefs and motives of others.*” This is in part the reasoning behind the Common Ground theory, which reasons that communication can be efficient due to “*mutual*

knowledge, mutual beliefs, and mutual assumptions” (Clark & Brennan, 1991; Clark, 1996). These assumptions might come from outside cultural references, understandings of power structures, or a shared referenced, but most often they are strongly based upon past interactions. Past interactions not only help develop stereotypes of actors (Simmel, 1910), but guide community norms for participation. The emergence of netiquette and the FAQ are two such examples of understandings that shape future interactions. Here historical context deals with the past history of semantic constructs created by a community or society at large.

The philosopher Paul Grice proposed four conversational maxims based upon the cooperative principal to make communication effective and efficient (Grice, 1957; Grice, 1969). Relevancy and (minimized) quantity are the basis for two of his maxims, but how is one able to achieve these goals when entering a new community? There has been no shortage of efforts in CMC to fill this gap under the pretext of shared Common Ground, although they have mostly been focused on synchronous communication (McCarthy et. al, 1991; Whittaker et. al, 1998; Kraut et al., 2002). Perhaps what is more encouraging to newcomers is the semantic compression of previous conversations. There have been only a few efforts in this worthwhile direction. Perhaps the most notable interface is Sack’s Conversation Map (2001), which integrates categories of discussion into what he calls Discourse Diagrams. Donath (2006) also attempted to paint a map of discussion, further contextualizing the semantic highlights with temporal and social relationships. Viégas’ Themail (2005) also visualizes past discussion temporally, but uses term-frequency based weighting of the raw words rather than creating higher-level semantic categories. While not explicitly about conversation, Galloway et. al’s StarryNight (1999) provides historical perspective by popularity within Rhizome.org’s database in a semantically linked network.

DIALECTICS

Dialectics guide what types of statements we would think to make, impressions we would receive, and given knowledge of our interlocutors, govern the strategies we take (Kunda, 1999). Dialectical opposition is most present when discussions are sided, such as political debates online. But dialectics also inform our realities, where we present assumptions as if there were no other perspective. Making CMC aware of dialectical oppositions can structure debate, improve hyperlinking to outside controversies, and expose points of view.

It is important to keep in mind how traditional arguments are constructed when building a CMC system in support of debate. According to Burleson (1992), the basic characteristics of an argument are: 1) the existence of an assertion construed as a claim, 2) an organization structure around the defense of the claim, and 3) an inferential leap in the movement from support to assertion. Because these segments may be individually considered contentious, we can improve

the navigation of a large set of messages by segmenting or hyperlinking aspects of a user's argument with related objects.

Intel Research has released a Firefox Extension called Dispute Finder that is designed to recognize and annotate claims or contentious viewpoints (Ennals et al., 2009). While not specifically targeted for online discussion, Dispute Finder creates parallel and coordinated views on top of web pages. It structurally differentiates between evidence and disputed claims, and offers hyperlinked fragments of arguments to be strategically applied. Its usage creates an abstracted graph of claims and counterclaims across the web.

Other works simply seek to highlight differences in opinion as a native component of the interface. Goldberg et. al's Opinion Space (2009) is a new project that enables users to contribute a variety of perspectives, which are then mapped two-dimensionally using multi-dimensional scaling. By traversing through the map, it is hoped that users can better understand the diversity of viewpoints similar and dissimilar to their own. Kittur et al. (2007) demonstrate automated techniques for discovering controversies in Wikipedia based upon revision histories. Their visualization segments users into implicit networks based upon their edit history, which illuminates core groups of shared perspectives within Wikipedia.

3.3 Content aggregation and abstraction

Despite years of online communities and accompanying designs ebb and flow, the principal method of large-scale interaction online remains asynchronous textual communication⁴. This technique has its advantages, such as being simple and straightforward, easily archived, and persistent. All experiments in this thesis perform analysis on textual data using machine learning.

Despite its ubiquity, analyzing text remains very difficult. Despite decades of research and recent advancements, Artificial Intelligence (AI) and Natural Language Processing (NLP) have not been solved to the point of achieving AI-completeness (Shahaf & Amir, 2007). Various techniques allow targeted successes, but technology is not yet near the point where we can simply ask arbitrary questions of the data. These issues are compounded by inefficiencies inherent in textual CMC, including 1) the tendency to diverge in topic or otherwise become entangled (Smith et al., 2000) 2) the lack of social cues of the poster (Donath, 1998), 3) the lack of passive social cues (Kiesler et al., 1994; Reid, 1994), and 4) the fragmentation of audience and information flow (Adamic & Glance, 2005). Luckily much can be done about these problems by addressing the underlying design and interaction strategy of the medium. The emphasis on 1-bit "*Like*" signals by Facebook is one example of simplifying the machine learning problem through interface

⁴ As of this writing, photo-based communication is rising dramatically. From the massive amount of photos shared on Facebook to startups like Path and Color, the rising ubiquity of smart phones has reduced the barrier to link sensors of the physical world to the Internet.

change (McCarthy, 2010). They export the burden of assessing quality to the users rather than having to judge the material themselves.

Beyond *Like* buttons, there have been a few mainstream attempts at aggregating textual information about people. Perhaps the most popular have been word clouds (also known as tag clouds), which are weighted lists of n -grams that manipulate visual attributes such as font-size, color, and order to convey importance or emphasis (see Figure 4.4). They can be used to visualize documents (Viégas et al., 2007), metadata tags or folksonomies (Hearst & Rosner, 2008), or otherwise characterize the context-free frequency of terms. The words are meant to conjure a semantic gestalt or to serve as a loose and flat directory. They are typically employed due to fashion over function as they have questionable usability, and are primarily oriented as portraits of users (Hearst & Rosner, 2008).



Figure 4.4. A tag cloud self-portrait by Wordle user number 3796367.

As word clouds are used to demonstrate frequency of a term, they suffer from limitation of what can be inferred. Various attempts have been made at sub-clustering within inconsistent semantic usage in user-generated folksonomies (Lux et al., 2007), discontinuous term frequencies alone can perform poorly in information retrieval tasks let alone summarization for humans.

More sophisticated text analysis tools aim to do a better job at clustering the words or documents against some

metric, or finding a heuristic to classify documents or authors along a given set of dimensions. Popular heuristics and goals include sentiment analysis (Pang, Lee & Vaithyanathan, 2002), taste modeling (Liu, 2007), social influence and information flow (Cha et al., 2010), social dynamics modeling (Khan et al, 2002), topic or theme surfacing (Blei & Lafferty, 2009), gender assessment (Mukherjee & Liu, 2010), and subjectivity analysis (Weibe, 2000).

These tools can all be used to surface aspects about people; however, finding a way to communicate these abstractions can be difficult, especially since the data is transformed in mathematical terms rather than structurally or using human-inferred semantics. The experiments in this thesis represent attempts at characterizing a range of textual data, from tweets to biographies, using a variety of techniques. While we are heuristically limited by the state of the art, we can maximize current tools to answer new socially-focused questions.

3.4. Social media and the commercial Web's attempt at profiling

Most CMC media today, otherwise popularly known as *social media*, do not attempt to give top-down views of the interaction that lie therein. Instead, the focus has become on the latest real-time information in the *activity stream*. The latest thoughts, passed along information nuggets, and evidence of offline behavior in the form of photos make up the bulk of signals available. Given the focus on communicating with friends, this should not be surprising as the impressions have already been formed.

On Facebook, Myspace, Twitter, Bebo, datemyschool, LinkedIn, Buzz, Orkut, Delicious, and Evernote our lives are presented in reverse chronological order. Assuming one wants to keep up with the latest of any one person, the design facilitates the consumption of a small amount of recent information (see Figure 3.5). However this principal does not adequately address the problem of gaining an accurate picture of an unknown person, yet the data could be used for such a task. The latest links, comments, and opinions begin the process of identity and habitus alignment, but it cannot reveal larger trends and anomalies. One would need to sift through much data to feel confident in making such judgments, a very time consuming process. The same would follow for their friends to continue the personal estimation, each person unmarked in its connection and increasingly costly for the information seeker.

To deal with the overload of information, most have targeted the consumer of a personal network's firehouse of information. Flipboard (see Figure 3.6) and other personalized *social media news readers* attempt to provide entertaining experiences to give users a handle on their network. Yet this ego-centric approach is always held in private view. If we are to believe the premise of these applications can be successful -- that one's network acts as an appropriate filter of online information to meet the reader's preferences -- then we can imagine that the same filters could be made visible externally to provide insight into a person's life and environment. As "[w]hoever controls the media -- the images -- controls the culture" (Ginsberg in Albrecht, 1980), we can extrapolate much about an individual through the culture in which they participate. Advertisers rely on such information to target their ads; this thesis seeks to shine a light on the consumer possibilities.

Many data modelers are increasingly using low-cost *Like* signals to conjure a model of a persons interests and influences. While this is a good way to assess the quality of information that passes through a network, for consumers the grand sum needs to be put forth in a way that is legible to outsiders and conveys a wide range of a person's positions. This may require surfacing more of the linked content, performing classifications of topic matter, or making higher level assumptions from the *Liked* particles of thought. Little research has surfaced in this direction, and remains an open question.

Outside of advertising, industry has focused on addressing only a small number of questions about one's aggregated behavior. The main theme has been finding individuals of value, whether

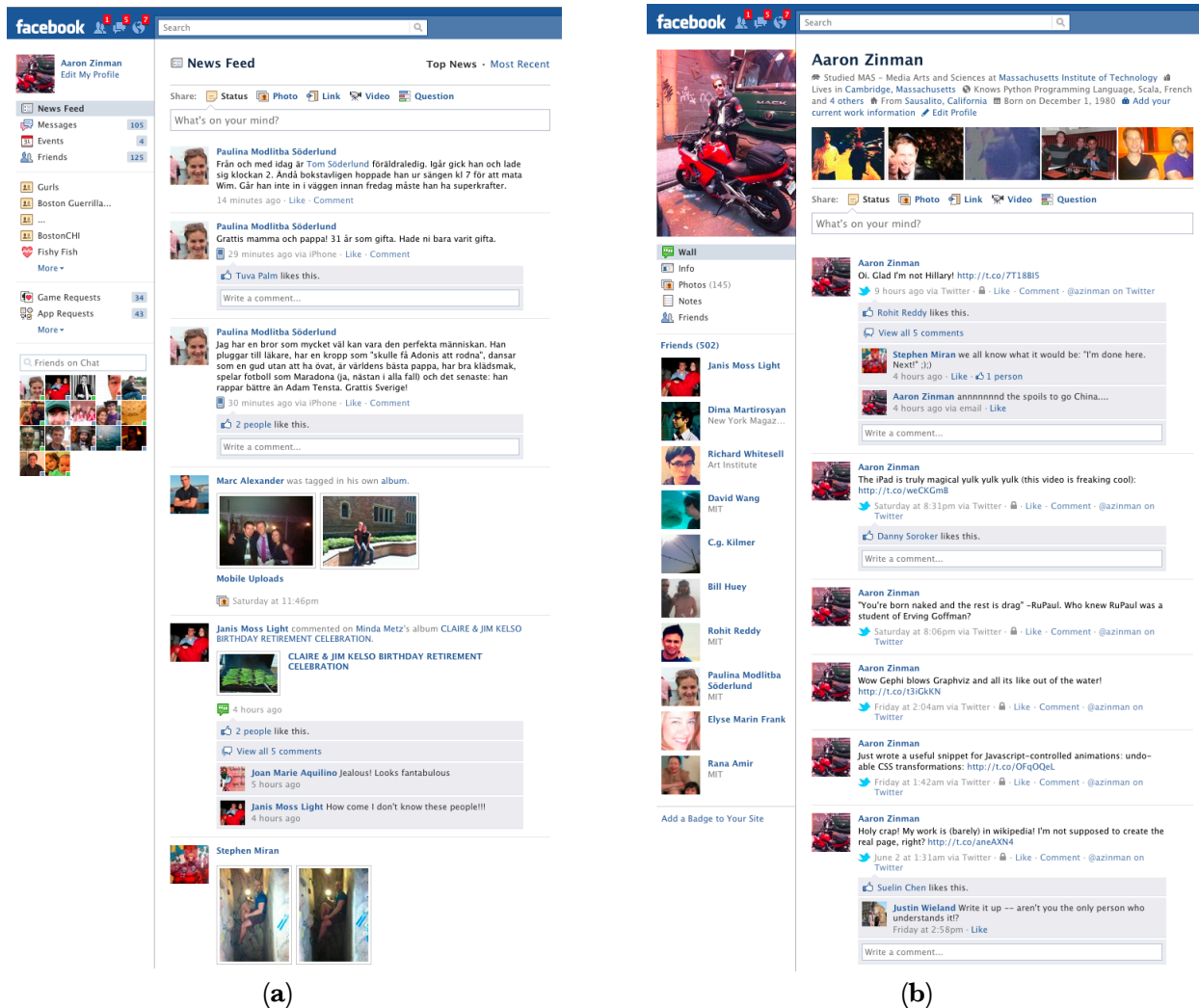


Figure 3.5. (a) The much imitated Facebook News Feed. Pulled June 6th, 2011. Life is shown in reverse chronological order. Individual profiles (b) are rendered similarly.

they are valuable from a military intelligence perspective (such as identifying terrorists), or those who have influence and reputation within social media. Klout is one of the more popular services on this theme, which computes a variety of statistics about the viral potential and other trends in Facebook and Twitter accounts to in turn compute a singular “*Klout score*” (See Figure 3.7). Klout further allows diving into the data to show which individuals are “*influencers*” and their “*influences*,” where the term influence is synonymous with reliably passing shared tweets and links. Klout also displays topics in which an individual has influence, which comes from automated algorithms finding associations behind the shared tweets.

Klout gives a good start towards answering these questions. Understanding the propensity for an individual to move information is an empirical indication of their stature. However, it can be difficult to interpret the significant of such information. The producers of content become

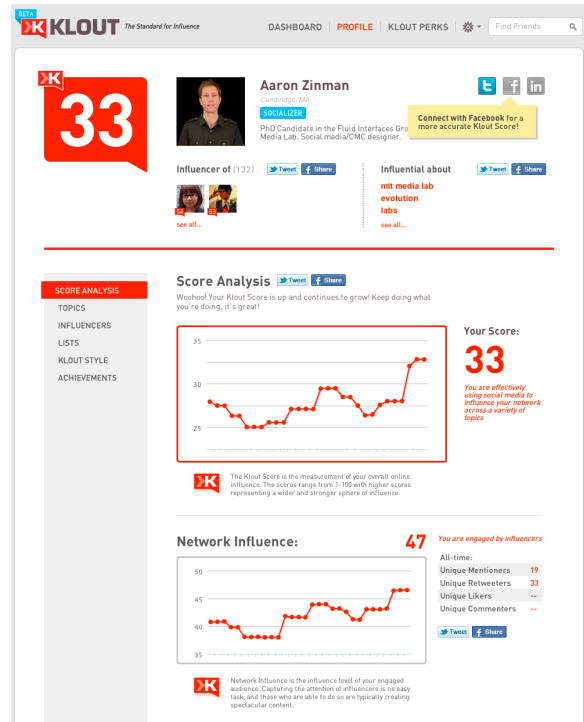
muddled with those who simply “pass the buck.” Furthermore ordinary individuals are not globally influential people, but that does not mean that there is no value in their existence online or otherwise. Industry needs to find a way to go beyond an individual score and into a deeper revelation of the history of an individual to gain a better insight into who they are in society and interpersonal expectations. Right now they are without control at the mercy of those who wish to compute their data into scores for exploitation and identification. However, these algorithms only know so much about those they compute; there needs to be a way for individuals to affect their representation. If not, the lack of control becomes akin to privacy concerns, where public data is being shown about individuals in a way they wish to restrict. To universally achieve this norm, we need more than technological innovation, we need social and legislative approaches towards data ownership and form (Lessig, 2006).

∞. Section summary

This section has examined existing works and methods of understanding social spaces and the people therein. Semantic versus non-semantic techniques have been explored, in addition a discussion of the role of visualization of social information. We conclude that much of these research directions have not made significant progress in everyday usage; industry has diverged on a different path to expose identity and preferences.



(a)



(b)

Figure 3.6. (a) Flipboard's social media news reader and information aggregator combines RSS feeds, Twitter, and Facebook data sources to provide an entertaining information consumption experience. (b) Klout combines this information and more to determine the reputation or “Klout” score of an individual.

4. EXPERIMENTS

In this chapter we discuss four experiments that illustrate ways of perceiving others using machine learning. Each experiment is situated in its own context, highlighting the myriad goals that can be achieved by synthesizing and visualizing digital traces. These works represent the journey of a researcher finding his way in making the hidden visible. At first, the methodology was more 1:1 in exposing the data surrounding an individual. With each new work, new algorithmic techniques were sought to expose higher-level attributes. While 1:1 visualizations can achieve perfect accuracy in the data they portray, this chapter argues that such an approach will quickly exhaust the available tools to satisfy the goals of both artist and observer. Instead, algorithmic approaches involving abstraction allow for more sophisticated data portraits by depicting previously inaccessible realms of the subject.

The first piece is called *Is Britney Spears Spam?*. It investigates the behaviors contained in the social network world, exposing predictable prototypes of socially-minded and promotionally-minded profiles. Moving into textual analysis, *Landscape of Words* seeks to provide a portal into very large communities by employing NLP to create a unique map. This map acts as a substrate to then further examine individuals while comparing them with their social network. *Landscape of Words* visualizes an individual within a single community, and *Personas* attempts to do so at Internet-scale. *Personas* surfaces characterizing-type statements about its users, and visualizes the machine trying to make sense of the data. It is a critique of a society that increasingly relies on data mining without understanding how it operates. Finally *Defuse* merges these goals by showcasing crowds and individuals together in a single interface. It demonstrates how communities and crowds break down in demographics that build on Bourdieu's notion of the habitus, and uses this as a point of navigation to deal with the increasing scale of crowds.

4.1 Experiment #1: Is Britney Spears Spam?

The question, "*Is Britney Spears Spam?*" is asked in reflection of an ever changing relationship between consumer uses of social media and the desire of advertisers. This work attempts to answer it by building a machine learning classifier to rate the perceived intention of approaching subjects on MySpace. It examines the network or structural level digital footprints of MySpace users to determine the limits of what can be prototyped without the difficulties of natural language processing. The results are intended to augment incoming friend requests to the observer with information to prototype what kind of user the subject is on the site. It can help users better understand the nature of a potential relationship by using machine learning to do what would otherwise be time consuming.

PROBLEM

People use social networking services (SNS) such as MySpace and Facebook both to stay in touch with people in their existing social network and to meet new people and expand those networks. Thus, communication with strangers or those you barely know are inherent to that world; they are constructed in part to enable unsolicited, yet friendly and welcome communication⁵.

This openness to messages from strangers also leaves users of these sites vulnerable to a growing quantity of unwelcome spam. Some would look familiar to any email user: ads for Viagra and breathy invitations to pornographic websites. Some are more ambiguous – is that “*friend request*” from an attractive stranger a genuine gesture from someone intrigued by your witty profile or is it phony façade that will lead to a torrent of advertising? Many sites, including MySpace, which was the subject of this analysis as it was the most popular at the time in 2007, reduce spam by requiring that in order to communicate freely with someone in the site one must be in their personal network. This is quite protective for users who are quite strict in chaperoning that network, never accepting anyone who is not well known to them. Yet for others, who enjoy the freedom of being able to make online acquaintances, there can be unpleasant repercussions from misjudging a requested connection. Spammers often pose as attractive young girls or other potentially intriguing characters in order to lure users to accept them into their network. Once a member, they flood the unwary user with a barrage of advertisements, or otherwise exploiting their network for nefarious purposes⁶.

As online social networking grows increasingly popular, so does the commercial use of these sites: people with something to promote, from pornographic websites to political candidates, are attracted to their huge audience and atmosphere of trust. For the participants, this means there is a growing need for technological assistance in sorting through advances from strangers. Such assistance is of course not new: this is what email filters do to protect us from vast quantities of spam. Yet in SNS, the problem is somewhat different: no longer black and white, there are numerous gradations in the desirability of contacts from strangers.

The definition of what constitutes spam in an SNS is often subjective. For example, one might receive a friend request from a celebrity such as Britney Spears. How much we love or hate Britney Spears might be independent of wanting to interact with her virtual persona. But unlike Viagra ads in e-mail, a non-trivial population actually does want to join the Britney Spears network. Thus, the role of the filter is not only to find the clearly unwelcome material, but to assist the user by highlighting and clarifying the most salient features of an unknown contact,

⁵ Some SNS services, such as Friendster or MySpace, are more suitable for meeting unknown people than others such as Facebook.

⁶ Using SNS as vectors for hacking attacks have been increasing, according to Symantec (Messmer, 2011)

making it easier and more efficient for the human user to determine whether they wish to accept the contact.

Vaughan-Nichols noted that spam, or unwanted messages, is almost impossible to define (Vaughan- Nichols, 2003). A penny stock ad is widely considered spam, but an advertisement from your bank might still be considered legitimate. Yet despite the gray area, spam has a clear enough definition such that e-mail providers Google and Yahoo will try to filter it for you starting from a master universal filter. Such master filters work for Google under the assumption unsolicited messages about medication, penny stocks, fake university degrees, and software discounts are universally undesirable. When Google misclassifies, we correct it by setting a binary spam flag. This approach towards e-mail spam is reasonable given a) we typically aren't contacted by many legitimate strangers, and b) we typically agree which messages should be marked spam. But what happens when both of these assumptions become invalid?

In SNS, it is no longer true that unsolicited likely means unwanted. SNS facilitate meaningful unsolicited communications, opening a large gray area for spam classification. Should spam filters take on the role of sorting through the full gamut of desirable and undesirable solicited communications? We think yes.

We postulate that for SNS, the redefinition of spam filters should start by focusing on the sender rather than the message. Content analysis might be enough to discover a Viagra ad, but often in SNS it is not enough. Requests to join a member's social network are contentless, only a link to the sender's profile. Thus we are required to judge the sender. If we are still only detecting the presence of select categories such as penny stocks or pornographic webcams, we can straightforwardly shift content analysis to the profile. But if we are rejecting a sender because they are a celebrity, we are rejecting a social prototype instead of the presence of select keywords. Without the capability to reasoning about people, we cannot adapt spam filtering to SNS; Britney Spears and Viagra are evaluated similarly.

Others have proposed that we can filter unwanted senders by injecting explicit or implicit trust values into the network (Golbeck, 2004; Levin, 1998; Kamvar, 2003). While providing a viable statistic when reliable, such systems only work well in the scope of friend-of-a-friend. As we compound multiple trust values to reach a node several hops away, our confidence in trust quickly diminishes as the nodes effectively become strangers (Donath, 2004). In SNS, it is exactly these strangers that we need to evaluate most.

Trust metrics are also problematic in that their definition is often one-dimensional. A single value cannot take into account how context changes the relationship between the same two members. For example, we might trust a friend not to purposely send us a virus, but we may not trust them not to send us marketing information about their company.

To reason about people more holistically, we need information to judge. Therefore, it is reasonable to assume network history is a prerequisite for a new class of spam filters to emerge. It is not without precedent; people have looked to construct social networks of private e-mail archives to identify spam (Boykin, 2005). But e-mail archives are private and incomplete. SNS already provide a rich source of information in their network structure and history of past actions and choices. Any public profile (which we calculate to be 78.7%) shows a user's entire social network within that site, personal information, and messages authored for the user by their network. The better and more complete the information, the more accurately we can judge the person.

Ultimately a person-oriented reasoning engine is needed to interpret the available information and present the results. This requires moving away from the popular e-mail solution of a binary spam flag; the ranges of unsolicited senders in SNS are more graded than e-mail. Therefore we seek a richer representation of people and content so that users and the filter can judge a broader segment of the social network. While this work is focused on a new breed of spam, the larger ramifications are to identify prototypical behavior at a structural and network level.

Creating a meaningful representation for both human and machine is non-trivial. As humans, we judge people on higher-level social rules than *is_penny_stock*. Disambiguating a fun attractive guy from a creepy attractive guy can be difficult even for humans, let alone a machine.

We believe a good way to represent senders is through prototypes. For example, a low-level prototype might be "someone who sends more movie clips to their friends than they receive," or "someone with little public information available." How we prototype users depends on our goals. If we want to reject Britney Spears, is it because she is a celebrity⁷ or is it because she unidirectionally broadcasts a lot of generic information? What we can identify as a prototype is strongly influenced by the features we can extract. It is much easier to measure public communication than it is to identify a celebrity, and might better reflect user preferences. Users might welcome Britney Spears as long as she spoke to you personally.

EXPERIMENT

User Characterization

We created a research prototype that characterizes users by their valence in two independent dimensions: sociability and promotion. We evaluate sociability by the availability of information of social nature. A large number of personal comments, graphical customization, and other

⁷ In SNS like MySpace, mainstream promotional entities are creating profiles and often joining as many social networks as possible. Their connections are often used as a marketing opportunity to open a one-way communications channel without consideration for the recipients concerns.

pieces of social human activity yield a higher score. Promotion is evaluated by the amount of information meant to influence others, whether its political beliefs or quite frequently information of a commercial nature. Typical e-mail spam would rate high in promotion, but low in sociability. They are trying to influence you, but there is no social dialogue. A local rock band, on the other hand, score might score high in both dimensions if they actively communicate with their fan base.

Note that sociability does not generically refer to the amount of information or content available. For example, we found that it is normal in MySpace to post a “thank you” to a member’s public message bulletin that just added you to their social network. Surprisingly, this happens frequently on profiles that have no intrinsic social value, such those that only promote a pornographic webcam. We consider such messages to be somewhat sociable, but without additional personal messages the score would be very low. On the other hand, Britney Spears could score high despite being a commercial entity in the presence of personalized communications with her fan base.

Rating a profile for promotion often requires a value judgment of the content. Are activists who speak out to their social network promotional? Humans can usually make this judgment, but it is difficult for a machine. We are interested in finding qualities of network usage that are harbingers of promotional intent. We suspect normal social human usage of SNS will have different character traits from solely promotional usages.

PROTOTYPES TO CHARACTERIZE

We previously mentioned using prototypes as a framework to allow users to express what they believe constitutes spam. We chose sociability and promotion because we believe the quadrants of their intersection represent four useful prototypes of users:

Prototype 1: Low sociability and low promotion. This user might be a new member to the site, or might be a low-effort spammer who doesn’t care about posing as something real. Without information to judge, we disregard such members from input to the classifier.

Prototype 2: Low sociability and high promotion. This is typical of a promotional entity using SNS as a marketing opportunity. They only broadcast generic information to the entire network, often trying to join as many networks as possible. Examples include Britney Spears, a Viagra ad, and a pornographic webcam.

Prototype 3: High sociability and low promotion. Such a rating is indicative of normal social humans. They connect and communicate with their social network on a personal level. They constitute the majority of active SNS users.

Prototype 4: High sociability and high promotion. Unlike the generic marketing approach, these promotional entities also engage with their network. Often small-scale media producers (local bands, YouTube directors) use SNS to connect with their audience, fitting this characterization.

DATA COLLECTION

Paulina Söderlund and I conducted an initial investigation to see if standard machine learning techniques could predict the classification of MySpace profiles in sociability and promotion using features specific to MySpace and its culture. We chose MySpace because it is the largest SNS and increasingly has become home to a wide range of promotional activity. Arguably, it was the promotional activity of bands that in fact made MySpace popular. However it now suffers from traditional spam and increasingly ambiguous intentions from large commercial entities.

We tried to capture a spread of such intentions by picking MySpace profiles at random, then rating them from one-to-five in sociability and promotion, where a higher number means a higher valence in that dimension. We will refer to scores by their variables s and p .

We only entered profiles into our dataset where $s > 1$ or $p > 1$ to only process profiles with information to judge. By the thousandth profile, only 11% of our database had $p > 1$; the majority of those were bands. We know that the number of promotional profiles is increasing, but our data suggests MySpace still has far more social-oriented content than non-social. Therefore we focused on growing our promotional dataset specifically until we reached 400 profiles where $p > 1$.

The 400 $p > 1$ profiles balanced against 400 profiles of $p = 1$ for the classifier. If we judged using the real-world distribution, a random guess of $p = 1$ would be correct 89% of the time. Given that we don't know if any of our features (to be explained) are meaningful, or if our dimensions are learnable, 90% accuracy is too close to a goal score. Therefore, we opted to balance the two sets. However, the 400 $p = 1$ profiles were selected such that they maintained the same distribution as the larger data set. Table 4.2 shows the breakdown.

After obtaining the contents of the profile and their ratings, we also collected the profiles of each person's "*top friends*." Top friends are differentiated by being explicitly featured in a subset of the immediate social network on the main profile page. This is interpreted in the culture of MySpace as showing one's "*best friends*" (boyd, 2006). We did not include the full graph not only to limit scope, but also because we hypothesize network-statistics influenced by meaningful social processes will highlight normal social humans.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$p = 1$	--	93	99	85	123
$p = 2$	1	5	7	16	60
$p = 3$	3	2	4	3	6
$p = 4$	46	17	5	3	5
$p = 5$	183	54	11	4	5

Table 4.2: Distribution of Profiles by Sociability and Promotion Ratings

Feature Extraction

Determining the best features for classifying people into our prototypes is non-trivial when we need to separate generic from personal. We hypothesize that network-focused metrics are among the most important statistics to distinguish people in SNS. Boykin and Roychowdhury showed e-mail spammers can be well identified as someone who has

many edges but few wedges (Boykin, 2005). Wedges arise from shared communities and geography, something spammers don't have. Yet a clustering coefficient cannot tell us that a user like Tila Tequila (Caplan, 2006), who tries to speak to as many strangers as possible, is not spam. Instead, network statistics can show that communication between Tila Tequila and her fan base is bi-directional, and that her users continue to propagate her media.

We selected our features by thinking broadly about how people use MySpace. This includes information available on the user profile, as well as the comments written on their top friends profiles. Our feature choice reflects social trends on the site, such as the use of detectable third-party content oriented towards MySpace profiles. Table 4.1 shows a hierarchy of our egocentric features, where “*topn*” refers a subject’s top friends. When we say “percent subject’s comments’ hrefs are unique,” we are looking for links within our entire data set to the same Internet address as a user has posted in their comments. Thus, it is possible many profiles in all of MySpace link to the same place, but we were unable to capture that. As a result, some of our features are inherently unreliable in our current configuration.

Network/Comment Based

percent comments from top n
percent top n comments from subject
percent subject's comments' images are unique
percent subject's comments' hrefs are unique
percent subject's comments' in top n hrefs are unique
percent subject's comments' in top n images are unique
average number posters use same images in subject's comments in top n
average number posters use same images in subject's comments
average number posters use same hrefs in subject's comments
average number posters use same hrefs in subject's comments in top n
number comments on top n
total number images in comments
total number hrefs in comments
total number images in comments to top n
total number hrefs in comments to top n
percent subject's comments have images
percent subject's comments have hrefs
percent subject's comments in top n have hrefs
percent subject's comments in top n have images
number independent images in comments
number independent hrefs in comments
number independent images in comments to top n
number independent hrefs in comments to top n

User/Profile Based

number friends
number youtube movies
number details
number comments
number thanks
number survey
number of 'I'
number of 'you'
missing picture
mp3 player present
static url available
has school section
has blurbs
is page personalized
has networking section
has company section
blog entries

Table 4.1. Features extracted for a subject by category using shorthand notation. The left-hand column represents digital footprints across the network using the socially-meaningful *top n* friends of the subject. *href* is a hyperlink to a given URL. The right-hand column shows the social signals extracted from the profile of the subject.

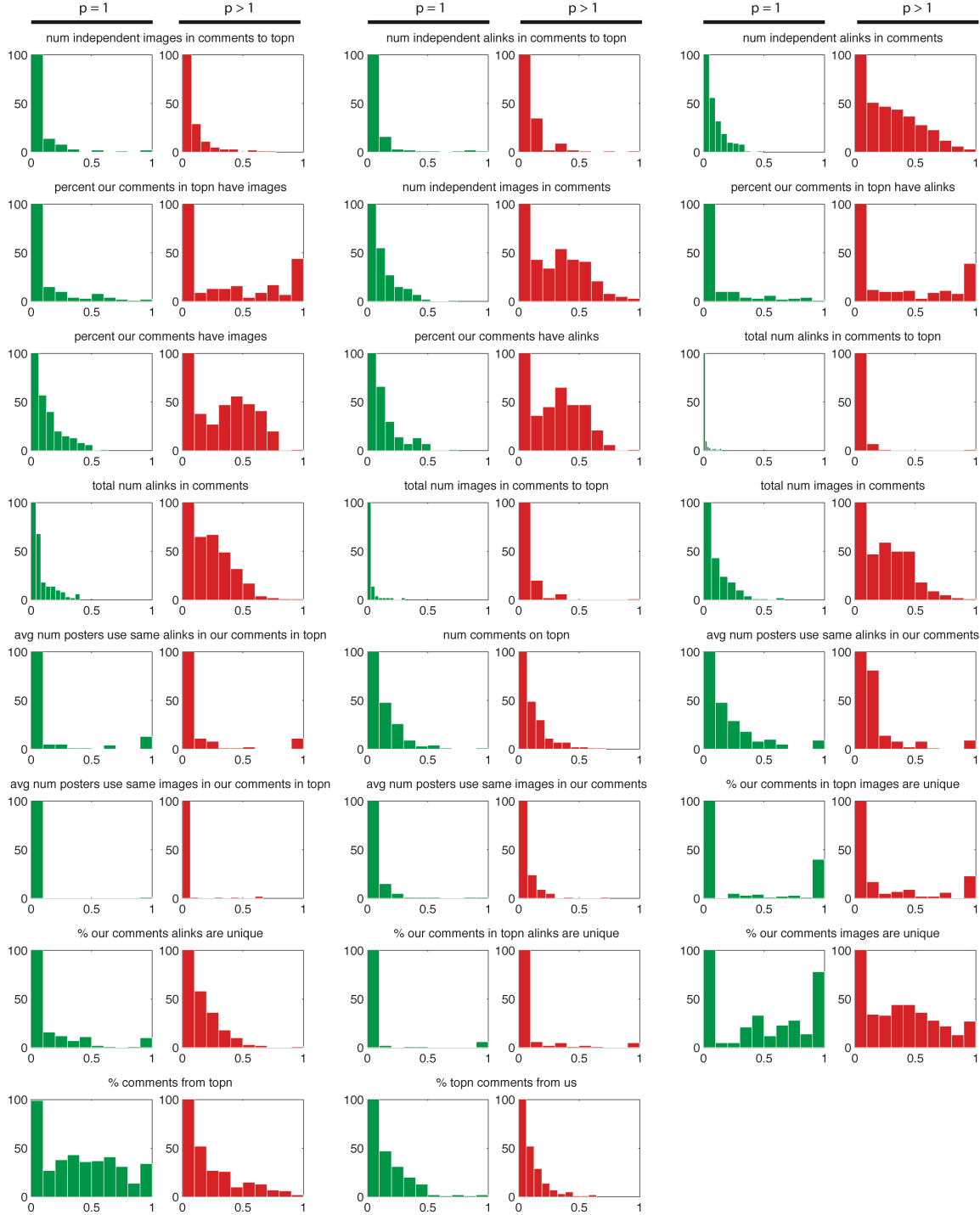


Figure 4.1. A histogram showing the results of network-based classification. Each data point is split into two columns depending on their promotional score, separating promotional entities (right and red) from the non-promotional (left and green). Note for the network features we purposely cut the graph off after 100 on the y-axis so as to visually concentrate the reader on the important details of the distribution while making a small graph size.

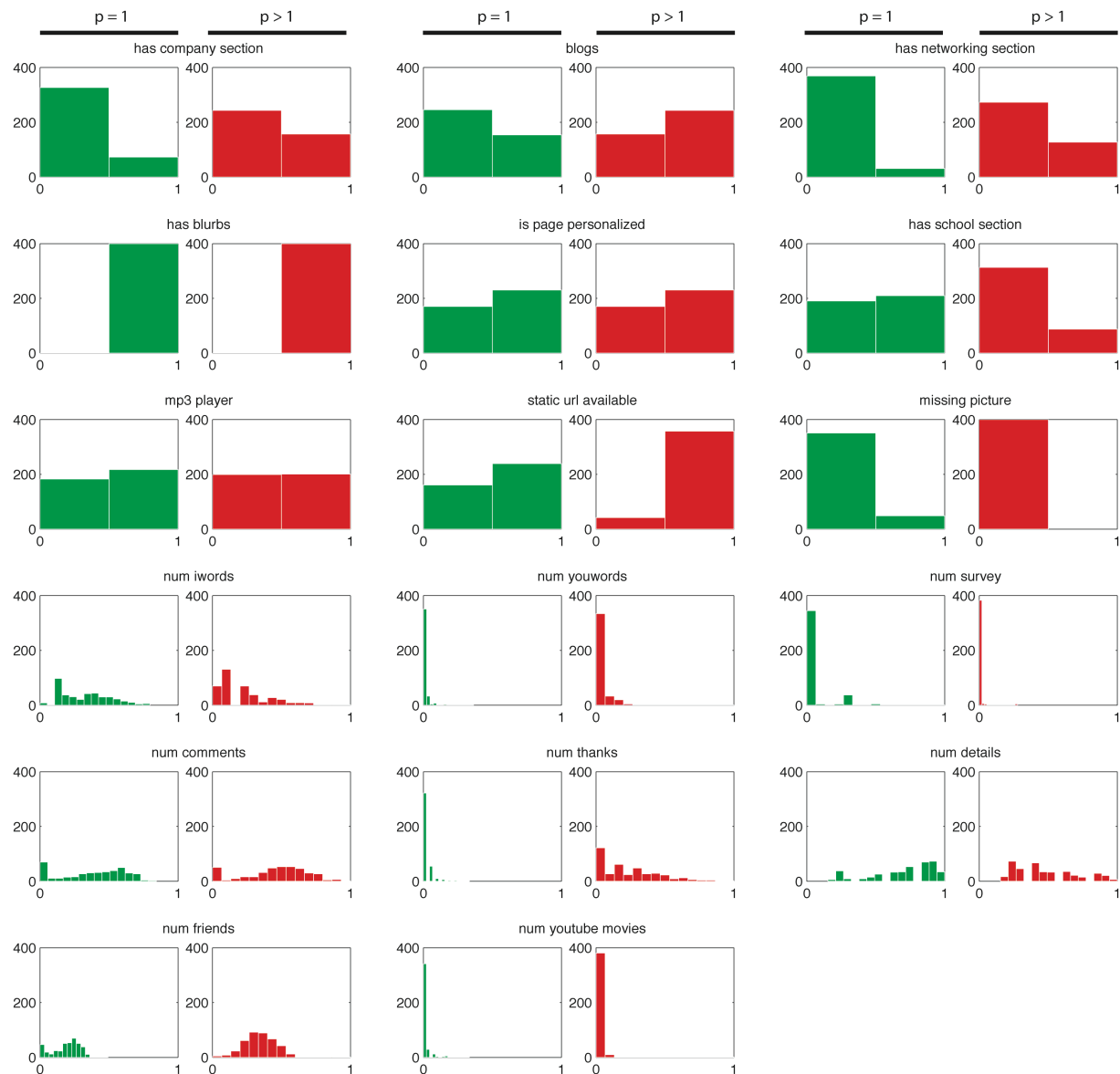


Figure 4.2. A histogram showing the results of profile-based features. Like Figure 4.1, each data point is put into one of two potential columns depending on its promotional score, separating promotional entities (right and red) from the non-promotional (left and green).

We normalized every feature from 0 to 1 so all dimensions could be compared linearly. Figures 4.1 and 4.2 show a histogram of the feature distributions of promotional-oriented profiles and those with no promotion. Despite a large bin around 0, most features display normal or power-law distributions. It is interesting to note that for several of the features, such as “percent our comments have images,” the type of distribution changes depending if $p=1$ or $p>1$. Thus we have evidence promotional entities use the network differently than non-promotional ones. Note for the network features we purposely cut the graph off after 100, so as to concentrate on the larger distribution while maintaining a small graph size.

Machine Learning

As we did not know if our features or data was learnable, we choose to survey many types of algorithms to see if any were suitable for our problem. We used linear regression, k nearest-neighbors, back-propagation neural networks (with varying number of hidden units and layers), and naive Bayesian networks. Each was ran multiple times using permutations of the following feature sets: profile-based, network-based, and mixed.

Given 40 dimensions and only 800 data points (600 train, 200 test), we feared the curse of dimensionality. We approximated feature selection using Principal Component Analysis (PCA) to reduce our space. We varied the number of dimensions kept with every learning algorithm, from 1 to 40.

RESULTS

Here will only discuss the results of our neural networks and naïve Bayes experiments. Their scores were better or similar to our attempts with linear regression and k nearest-neighbors.

Our networks performed poorly in correctly classifying a profile in both dimensions simultaneously. The network did not do much better than 30-50% in any configuration, which is still better than random (see Table 4.2 for typical performance).

As we will later discuss, there was a large amount of subjectivity in the hand rating of the profiles. Due to time constraints, our hand rating only underwent a single pass per profile. Thus there is a high probability that another pass at the same profiles would result in the slightly different score, even from the same original reviewer. To handle this situation and get closer to how a human might expect to interact with a filtering agent, we created several new tests based upon a notion of thresholding. Our thresholding function seeks to correctly guess which side of given value (from two to five) a profile falls in a given dimension. For example, if our threshold is at three, and the data is actually one and we guess two, we would count that as correct because everything is on the same side of three. However, if we guessed three and the correct answer was two, our test would evaluate to false. The thresholding function reduces the subjectivity in our original ratings by fuzzing the scores while concentrating on a single dimension.

We created threshold tests to classify each dimension independently, in addition to a special “firewall” threshold that crosses both dimensions. Firewall is a special test that tries to represent the spirit behind current spam filtering, which is to block out non-social promotional entities. It is the same as the promotional threshold test except we also require $s > 1$ (the profile is at least somewhat social). The user then sets the maximum promotional value a profile may score and still be let through.

All of our threshold tests unsurprisingly performed significantly better than the exact requirement tests, showing that at least something useful could be extracted from our features and dataset. For the firewall test, our performance ranged from 90-93%, with the best performance at $t=4$.

Surprisingly, we found that reducing dimensionality using PCA did not improve performance: much of the reduction actually gets performed by the trained network. This was also evident by the fact that fewer hidden nodes performed extremely well in our Neural Networks. Thus, we conclude the task may be inherently more linear or less multivariate than we previously assumed.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$p = 1$	--	0%	0%	0%	70.5%
$p = 2$	43.5%	0%	0%	28.6%	7.7%
$p = 3$	26.9%	0%	0%	0%	0%
$p = 4$	--	--	--	--	--
$p = 5$	55.6%	0%	0%	--	0%

Table 4.2: Distribution of Profiles by Sociability and Promotion Ratings

The best performance came from using both feature sets in a single layer neural network (Figure 4.3). However, this was only marginally better than using only profile-based features. We conclude that there is still value in including network usage statistics, but our profile-only features were good enough to get us most of the way there. The network-only tests fell between 78- 83% accuracy, much lower than

with the profile-based features. While this might seem discouraging when our goal is to use network-based features, we hypothesize that our preliminary feature set has much room for improvement by using more robust network statistics. For example, we did not include timestamps of comments in our features. The networks and comments of a “real” persona are built up organically over time, a process and resulting network and communication pattern that is difficult for spammers to mimic.

Performance of *promotional* threshold across the best performing network configuration for each model

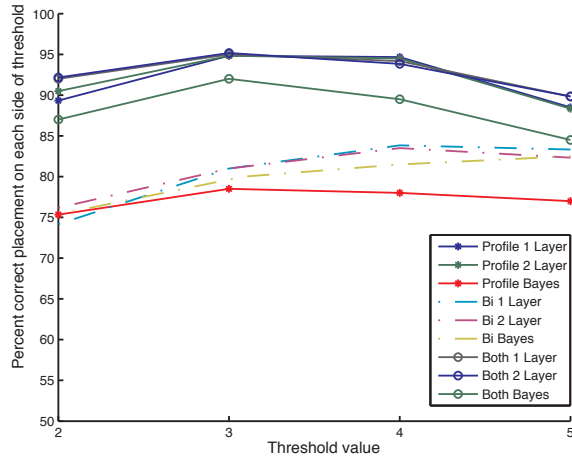
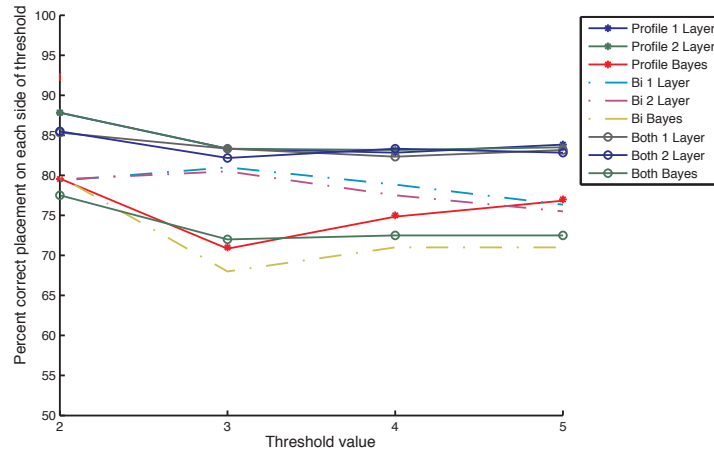
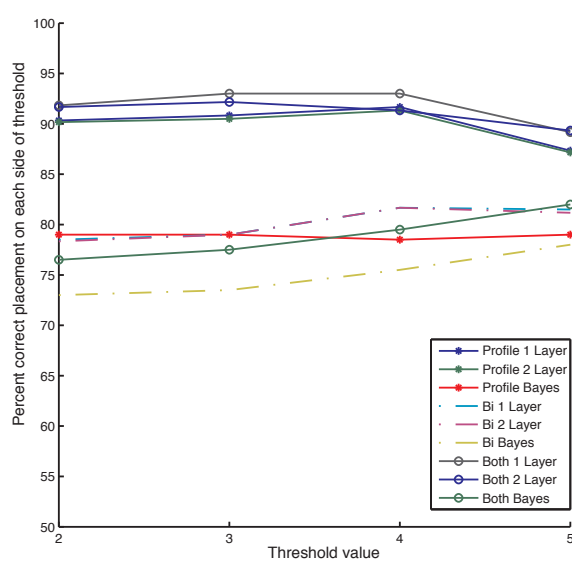


Figure 4.3. Graphs showing the best performance of each classifier permutation for each of the three threshold-based tests.

Performance of *social* threshold across the best performing network configuration for each model



Performance of *firewall* threshold across the best performing network configuration for each model



Profile scoped features will have a limited time that they can be considered useful in the spam/anti-spam arms race. We currently see a large increase in e-mail of image-based spam, simply because it is more costly for modern filters to handle. While spammer techniques will always adapt around the current detection technology, we believe a network-centric approach is ultimately more robust.

DISCUSSION

We believe we have identified a promising conceptual scaffold to filter solicitations in SNS by using the concepts of prototypes and feature bundles. This approach can be used to power future data portraits. Although our preliminary results are less substantial than we had expected, we believe the flaw to be in the choice of analytical techniques rather than the underlying network-centric approach. A logical next step is to use more advanced techniques to analyze the network for the separate purposes of deception detection and human categorization. The remaining three experiments shifted the analysis from the network onto content to explore the analysis of more sociologically-grounded prototypes.

As previously mentioned, Boykin and Roychowdhury have shown the clustering coefficient of a generated social network to be useful in fighting email spam (Boykin & Roychowdhury, 2005). They first examine the headers of an individual's email archive to approximate the actual social graph, then using its network properties classify users into white and black lists. While their methods could only be applied 47% of the time due to algorithmic constraints, when applicable it works fantastically well. Clustering coefficients are a promising example that network properties can at least usefully distinguish normal human behavior from the purely deceptive and malicious. Kimura et al. (2005) showed a similar technique can work well for search engine spam within trackback networks. As we have already discussed, it remains an open question which network properties are appropriate given the changing subjective goals of observers and the typical usage properties of a given site. Clustering coefficients are only useful if the culture of the network supports it.

We believe more research in passively generated statistics of SNS usage can get us much of the way there. Usage is influenced by preexisting social conditions; we bring our cultural norms, communities, schools, geography, and friends into the networks we use. Sometimes local properties like geography can be a stronger force to grow the network than the network itself (Marmaros & Sacerdote, 2006). Some patterns, such as temporal rhythms (Golder, Wilkinson & Huberman, 2006), function well as markers of average human activity. More social science research into SNS is needed to distinguish the different types of users and cultures within a given network (boyd, 2006; Fiore, 2004; Golder, Wilkinson & Huber, 2006). Such work is invaluable when algorithmically applied to detect humans and the various categories within them.

The features we choose directly impact what we can predict and what we can show the observer. If the desired categorization is too ambiguous or high-level, even the best classifier engine is likely to perform poorly. We chose sociability because we believed it matched the raison d'être of SNS; promotion reflected the growing misuse of SNS. We now realize they were harder to achieve because their machine evaluation requires value judgments difficult for even humans to make. For example, how sociable is (MySpace commercial entity) Britney Spears? Do personal responses from separate public relations interns constitute sociability? Do her "friends" need to actually know her in real life? As we further dive into the analysis of profiles, we uncover even deeper philosophical questions that challenge our assumptions and expectations of a virtual identity. Must only one mind represent an entity? Does "it" need to be human? Does it need to be just one human, or can it be two humans and a dog? What if it is clearly a human but is primarily about their business? Such questions highlight how arbitrary our current definitions might be as computer scientists when proposing generic anti-spam solutions.

Until we have reliable agents using a fine-tuned subjective cognitive model of the observer, future work should examine how the network features could power additional prototypes. For example, the average ratio of messages sent to received might be enough for most people to filter a majority of profiles to their liking at a first approximation. This works because by itself it can be understood as meaningful social statistic: "Britney Spears" can be worthwhile as long as the subject usually converses back.

SUMMARY

We have argued that sorting approaching strangers needs to be more nuanced than the black and white, spam or not spam classification typical of most email analysis tools and social networks. We need to be able to classify a range of potential subjects to assist observers of varying interests and tolerances in deciding which unknown contacts to accept and which to discard.

We attempted to do so by creating a model that could rate subjects in the dimensions of sociability and promotion. However, we quickly found that doing so requires placing a value judgment. When we, humans, were hand rating profiles to generate our data set, we often disagreed about what score a particular subject should take. For example, are political activists promotional, or is that only reserved for those selling something? If it is difficult for humans to agree on a particular rating due to subjectivity, how can we expect machines to perform the same tasks for us?

Only the observer can decide if Britney Spears is spam. Yet the design of SNS and their associated services can speed this evaluation through digestion and presentation of information that would otherwise be hidden. Facebook has already begun the practice of publicly consolidating and aggregating activity of its users for consumption in its popular *News Feed*

feature. However, it functions as a social radar at a literal level rather than a predictor of potential activity. If we expect the premise of social networking to continue to expand, new interfaces will have to be built that highlight any past behavior indicative of future behavior. Without advanced artificial intelligence, we presently advocate the presentation of facts without using subjective language or categorization.

We are confident that harbingers of promotional intent can come from the analysis of network usage qualities. Regardless of our subjective follies, our histograms have shown at minimum that ordinary people and promotional entities have some differing character traits in network usage. At first this may not seem surprising, but the differing traits go beyond “how many people they attempt to befriend or contact.” Clustering coefficients, gradients of bi-directionality in communication, and media sharing practices all give us insight into the behavior of entities that may be otherwise unreadable or too easily falsified. Future combinations of natural language processing with social network analysis have the potential to give an accurate prediction of what to expect from an unknown entity. It should be principally supported by examining an entity’s role within the context of their friends and the culture across the entire site.

As John Keats famously wrote, “*Beauty is truth, truth beauty, that is all Ye know on earth, and all ye need to know.*” In the vulnerable world of SNS, the truth may be ugly, but being able to reliably digest and present usage facts may be their only hope to preserve utility and curb chaos.

4.2 Experiment #2: Landscape of Words

On the Internet, new network-oriented communication services come and go. Some are very specific in their intended usages, such as the recent academia.edu which aims to allow researchers to share papers, reviews, and stay up to date within a field. Others are more flexible, like Quora or 4chan. Depending on how much structure has been provided to guide the intended usages, it takes time for social norms of a given network to develop in conjunction with the initial community. During these critical early stages, it can be difficult to assess the utility and typical usage patterns of a given medium. It would be beneficial to service providers and users equally if we should shine a light into the diversity of norms as they begin to establish themselves. Reducing the friction to understand bubbling usage patterns could help accelerate adoption or dismissal. *Landscape of Words* was an attempt to examine Twitter at its early stages by shining a light on cultural practices. It is a data portrait that uses *Latent Dirichlet Allocation* to find word patterns called *topics* within the Twitter corpus, and uses *Multi-Dimensional Scaling* to help visually project the model onto a map using a geographical metaphor for the observer. It demonstrates to the potential for such approaches to provide a scaffolding and zeitgeist for any emerging or existing medium. This work was done for the NSF Visualization contest in collaboration with colleague Alex Dragulescu.

PROBLEM

Twitter is a social networking service that has experienced explosive growth in its short lifespan. The service allows users to post messages to their directly accessible public profile. These messages can also be aggregated across multiple users by “following” them, creating a decentralized broadcasting platform. These messages, or tweets, are very short: they are limited to 140 characters. The principal utility of the service comes from the combination of human-compressed messages, an agile publish-subscription, and the easy of information flow through the ability to “re-Tweet” a message to one’s own audience.

As of March 2011, one billion tweets are sent per week (Twitter, 2011). This rate is extremely impressive especially given the service was only started in 2006, and took over three years to reach the first billion. Around 2008 Twitter started exploding in popularity. As it began moving from tech-focused early adopters into the mainstream media, there was much of confusion as *raison d’être*. With a 140 character limit, the low-cost of sending a broadcast asynchronous message was met with confusion as to what to broadcast, given the norms for such a medium had not been fully established (Walther, 1992). Various trends emerged, from the seemingly banal reporting of everyday activities, to celebrity culture and the passing of information (Marwick & boyd, 2010). Because the culture surrounding short public messages had not yet developed, how was an ordinary person to know why the hype should apply to them?

In 2008, one way of assessing Twitter’s usefulness was to simply look at the public tweet timeline. Presented in reverse chronological order, Twitter’s homepage displayed a random assortment of unconnected tweets (see Figure 4.5). Because these Tweets come in randomly from multiple languages, audiences, and contexts, it can still be difficult to assess the primary affordance of the medium. Adapted norms such as hashtags and at symbols appear without legend or explanation, only complicating the ability digest the tweets that do appear.

However, the heart of the problem lies at the form of presentation: a long flat linear list at granular level. Without a top-down perspective into the data, combing through the items in the list can give some insight only if the usage is homogenous and intuitive. Unfortunately, for Twitter and most popular services, this is not as much the case. Twitter in particular is susceptible to trends at various temporal frequencies; in fact, usage of Twitter is so trendy that it has prompted Twitter and offshoots to develop algorithms to capture those trends. Looking at a given point in the timeline does not easily afford understanding if a given trend is present, and if so, its temporal granularity. The typical format of a list is for displayed events to be so recent that going back two months is not feasible, let alone understanding the past several years.

Furthermore, the list itself is so visually focused at the Tweet level that assessing the range and power of audience is ill-afforded. The users of Twitter dictate its usage, yet the demographics are



Figure 4.5. The public Twitter timeline shows a random assortment of the latest Tweets from across the globe.

characterize the social dynamics such as PeopleGarden (Xiong & Donath, 1999), Seascape and Volcano (Lam & Donath, 2005), and Loom2 (boyd et al, 2002), they do not allow summarization of a service at a content-level. Instead, theme-extraction typically takes on metaphorical tones in the visual domain. ThemeRiver uses a river metaphor to describe the ebbs and flows of trends across time in a stacked graph (Havre et al., 2002). In Figure 4.7, ThemeRiver is used to visualize the lifespan of words associated with documents about Fidel Castro across a 40 year time span. The human-annotated interactive visualization allows one to zoom as needed to zero in on a particular event, or to dilate time to understand the larger trends. Incremental improvements to ThemeRiver have encompassed how it summarizes and displays its data (Liu et al., 2009), as well as improvements to the visualization's aesthetics

seemingly hidden. Ironically, the power of its network to transmit and pass on information is one of its greatest strengths and usage (boyd, Golder & Lotan, 2010). Issues surrounding class and race become paramount when trends started to appear that came from audiences culturally separated from the original community “owners.”

Landscape of Words sought to answer the question “*What does Twitter look like?*” Given the diversity of cultures, trends, and norms of a given site, it is even possible to create a data portrait at mass scale? If so, could it become a backplane for navigation within the site itself and thus providing context to a range of apertures when inspecting the site and its users?

INSPIRATION

There have been various attempts at massive aggregation and visualization of communities for navigational purposes.

While some systems attempt to

(Byron & Wattenberg, 2008). Conversation Maps (see Figure 4.6) attempts to find the themes for Usenet newsgroups and depict their interrelatedness on a connect graph, in addition to other views into the dataset (Sack, 2001). Finally most related work for Landscape of Words is ThemeScape (Wise et al, 1995), later commercialized by now defunct Cartia Inc. as a product called NewsMaps. ThemeScapes are generated by a proprietary text analysis engine called SPIRE, clustering documents by lexical commonalities into a 3D form that resemble mountains. While Landscape of Words builds on ThemeScapes, it was created without existing knowledge of ThemeScapes.

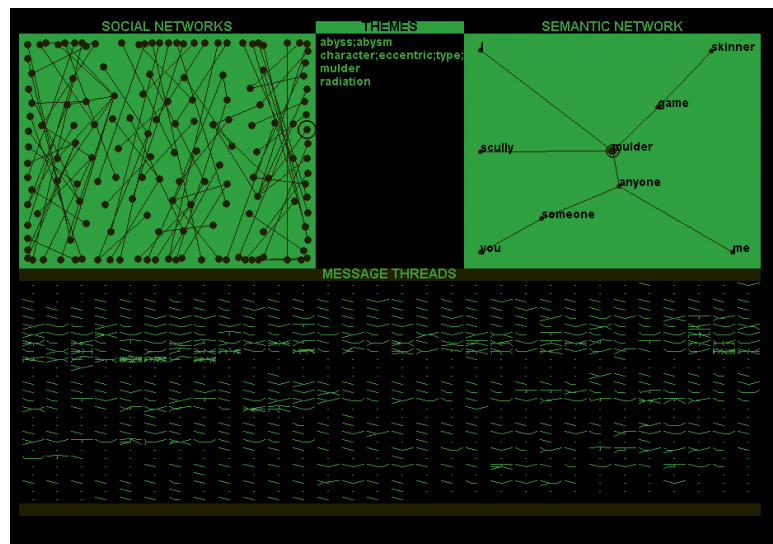


Figure 4.6. Sack's (2001) Conversation Maps. The visualization reveals multiple aspects of Usenet newsgroups, from simple structural mappings to more the abstract in theme extraction and semantic networks.

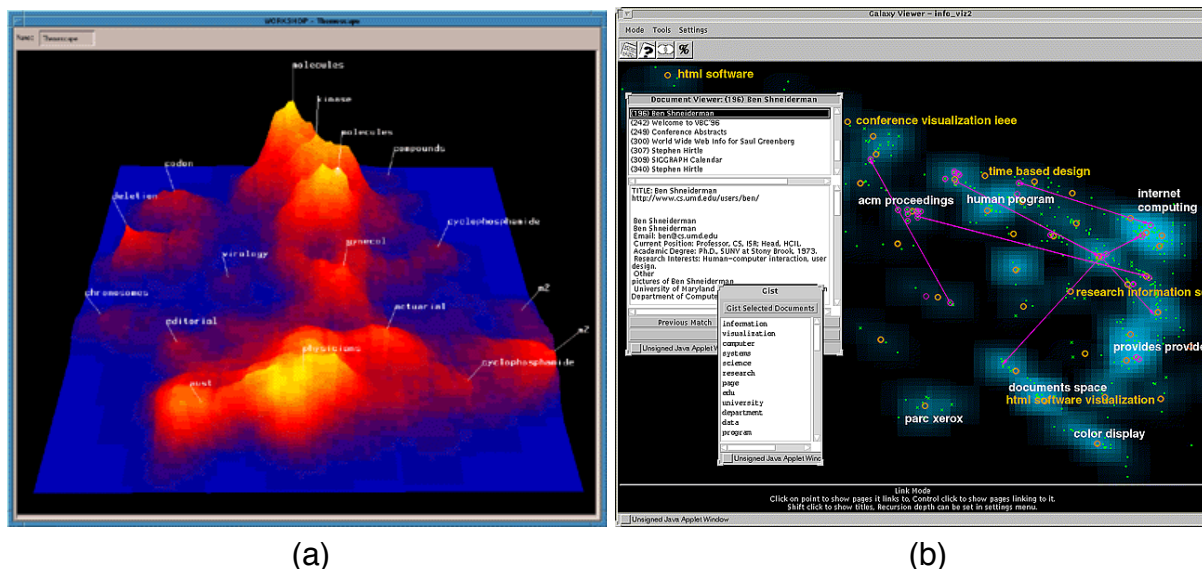


Figure 4.7. (a) A ThemeScape visualizes a large set of documents using a 3D landscape metaphor. Documents are clustered into mountains that are annotated by their common properties. (b) WebTheme uses a galaxy metaphor to show relatedness between document clusters in an interactive viewer.

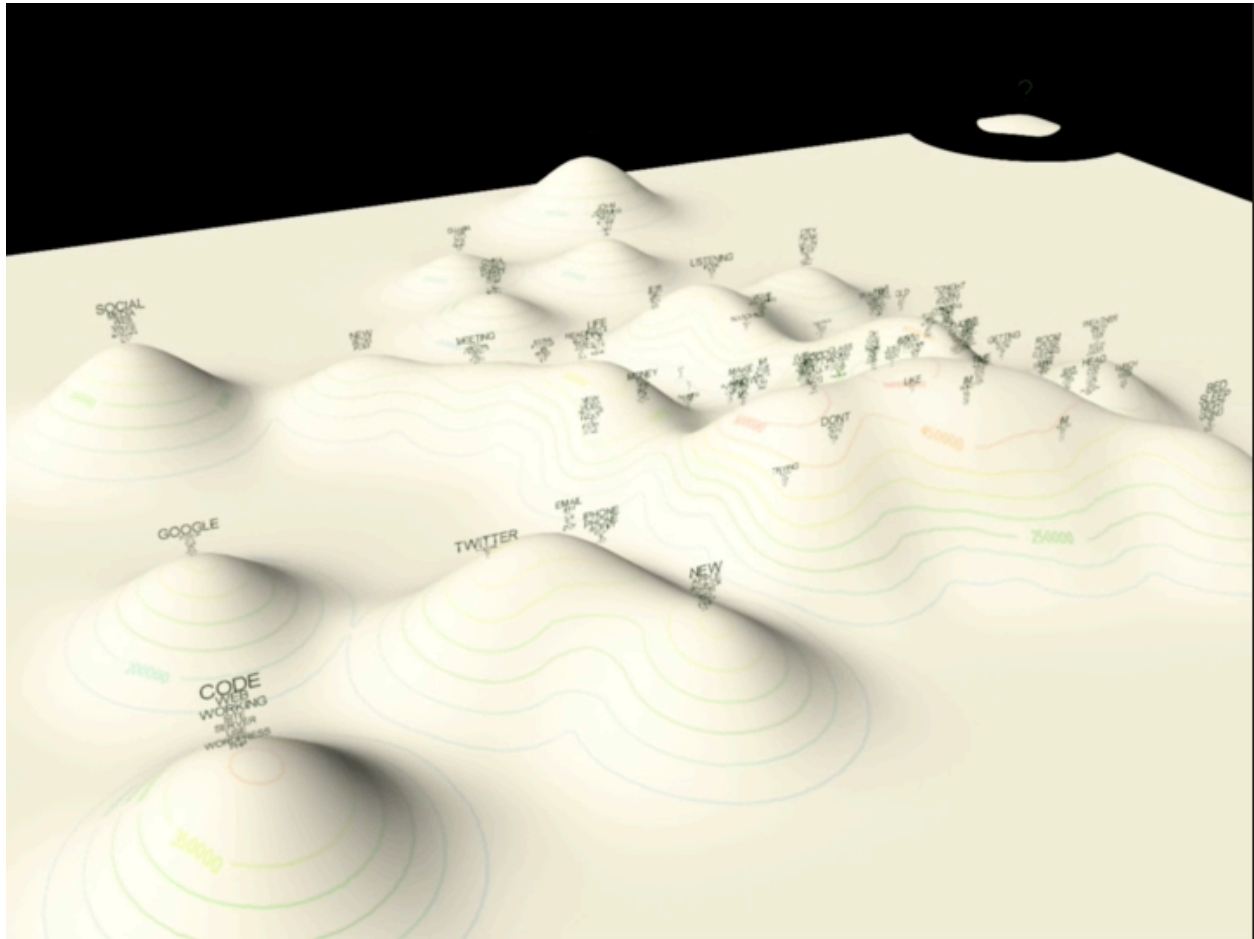


Figure 4.8. Screenshot of Landscape of Words. We see the topic mountains in their base state, where each bundle of words describes the topic by its most probable words.

DESIGN

Landscape of Words was designed around an algorithm called Latent Dirichlet Allocation (LDA), which is discussed in further detail in chapter 7. LDA is what is called a topic model, which are able to find themes or “*topics*” that emerge in large collections of text without requiring any preexisting knowledge. To visualize them, we chose an approach similar to ThemeScape using a mountain metaphor for these topic clusters. Each mountain represents a distinct topic, where its height is proportional to the number of times that topic is assigned to the Twitter corpus (see Figures 4.8 and 4.9). The most probable words for each topic are vertically arranged on the peak of each mountain, each sized in proportion to their probability of membership. The topic may be interactively explored by hovering over a mountain to increase the size and thus legibility of its top words. Mountains are placed near other mountains that are similar according to the latent semantic model. A topographic boundary contextualizes the mountain sizes, indicating the number of tweets that have fallen into a given topic. In the corner, a small offset mountain represents Tweets that did not fit well into any of the topics.

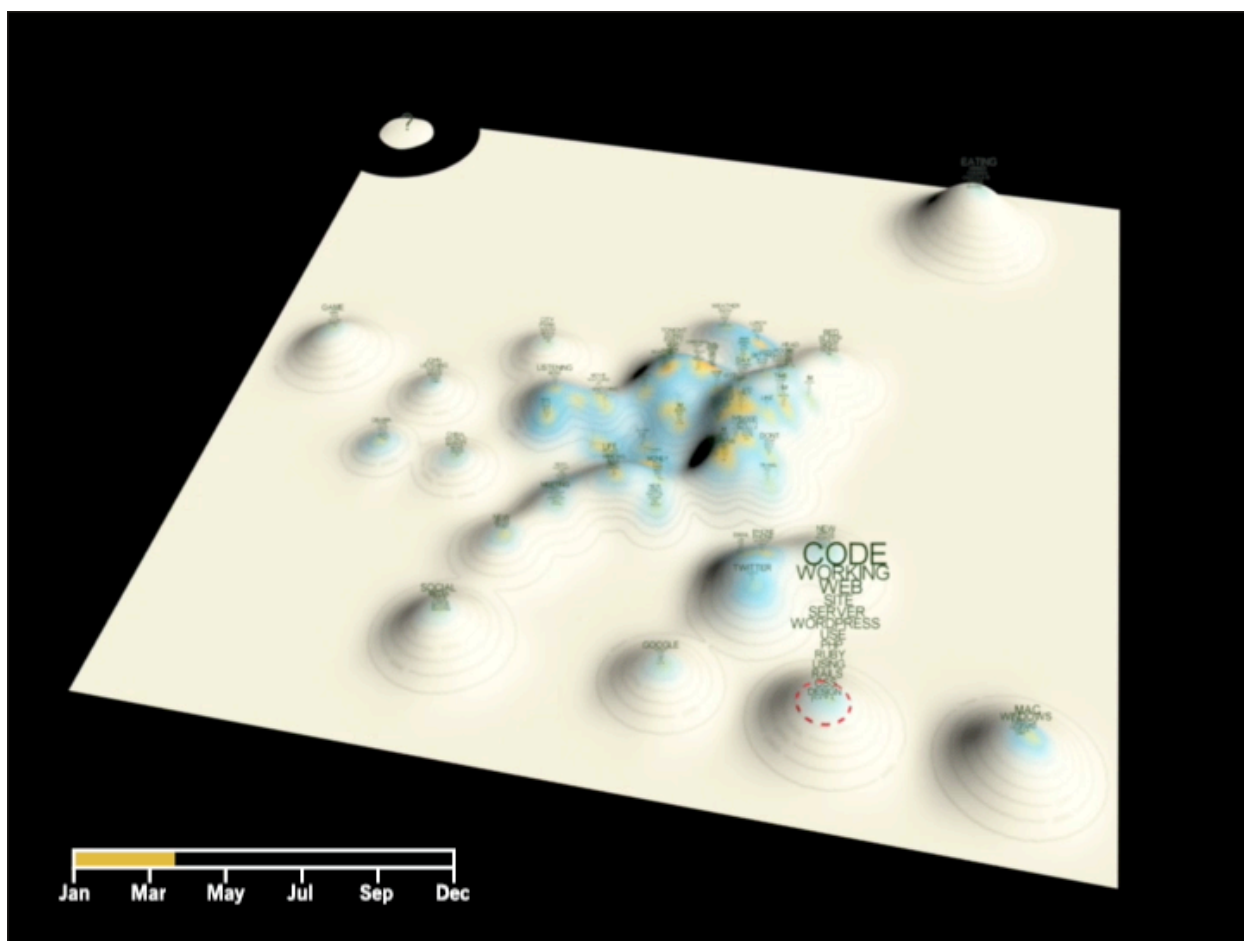
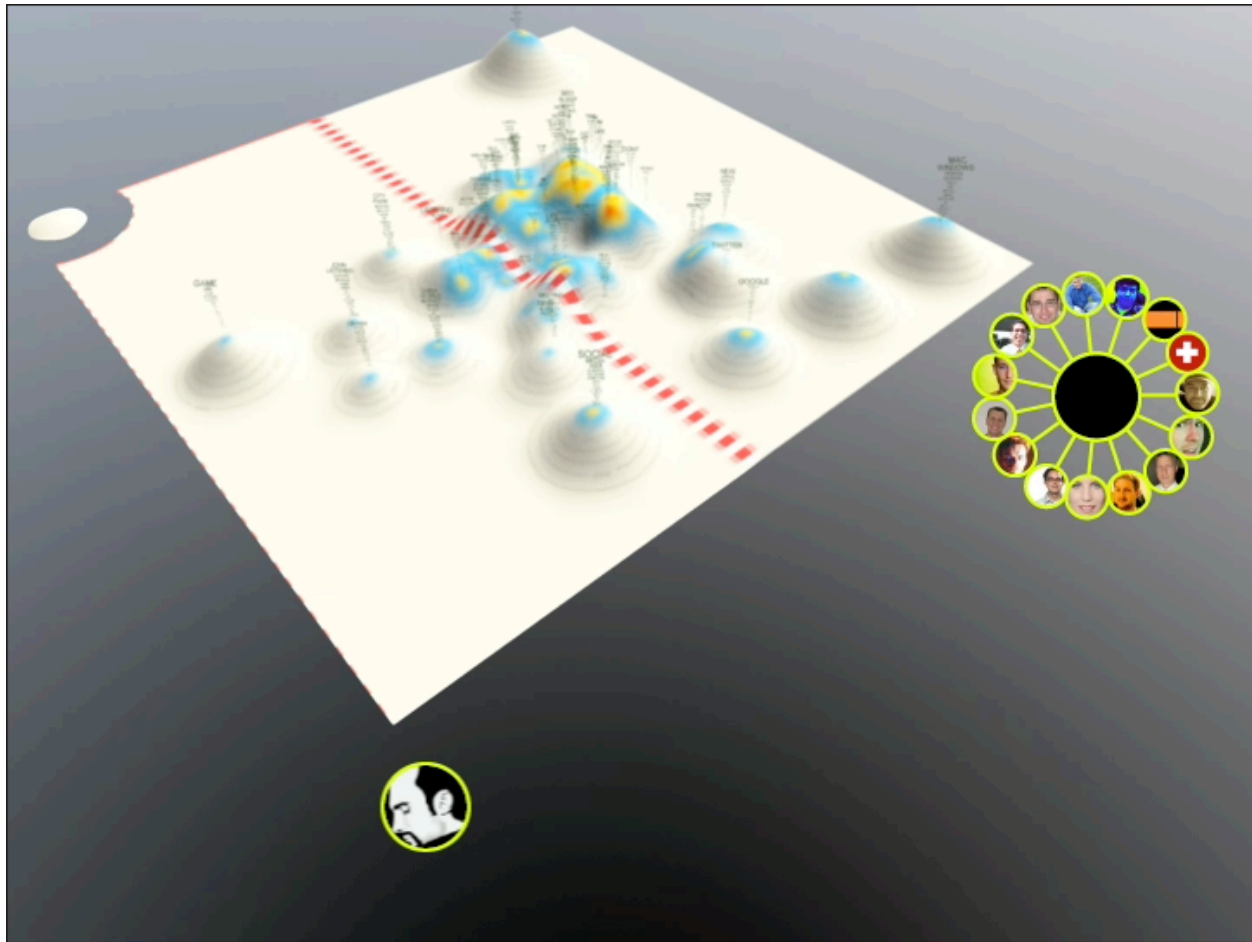


Figure 4.9. Screenshot of Landscape of Words. Here we see an interactive exploration of the mountains to zoom into the individual topics. A heat map is overlaid and animated to show the ebbs and flows of topic popularity across all of Twitter in 2008.

At this point Landscape of Words differs from ThemeScape in that a) it uses LDA to capture a wider variety of conceptual semantic and sociolinguistic groups than ThemeScape’s proprietary analysis engine could provide, b) it is able to scale to millions of documents rather than the 20,000 limit of ThemeScape⁸, and c) offers a place to put ill-fitting documents. We sought to then extend the visualization by using this topic mountain as a common substrate to overlay additional data.

As the landscape is built using all of Twitter as its basis, it serves as a common reference point that is relatively static and learnable by users. We may then overlay heat maps on top of the landscape to visualize the topic concentration of a given user, set of users, or by animating

⁸ The 20,000 figure comes from the original ThemeScape paper, published in 1995. IN-SPIRE, the commercial offering from the Pacific Northwest National Laboratory that is the modern day version only supports 200,000 documents.



through time, all of Twitter. It also goes beyond its original purpose in trying to visualize an LDA model to explain what Twitter is, and into the realm of condensing information about individuals and groups. Particularly relevant are the groups of users that follow an individual, or that they follow. Applying principals of homophily, it is useful to see how much divergence exists between an individual and who they find interesting. The visualization employs a dotted red line to act as a fader between the two heat maps, where the heat maps of one individual/grouping appear on one side of the line, and the heat maps of the other individual/grouping appear on the opposite. By sliding the red line back and forth, one can get a sense for the similarities and divergence between the two sets.

IMPLEMENTATION

To implement Landscape of Words for Twitter, scrapers were created that both pulled from the main tweet timeline as well as spiders that followed the following-graph per user that appeared on the timeline. Because in 2008 Twitter was much smaller, all of Twitter (the link graph and

timeline stopped producing unknown individuals) was able to be mirrored with approximately 10 million tweets.

Once mirrored, the data needed to be cleaned up and tokenized. We first performed language detection using a simple filter that first filtered on character set encoding, and then using a *trigram* model trained on NLTK-provided corpora. Next, we tokenized the document using our custom multi-stage pipeline approach built in Python called *Tokup*. Once non-English, garbage, and tweets less than four non-stop words were removed, a corpus of approximately 5 million remained.

While this thesis does not go into the details of Tokup, it represents a considerable effort (about 4 man-months) towards the tokenization of real world messy Internet-based text. Tokup was originally crafted to robustly parse Myspace profiles, which is how Table 4.1 was able to be created. As language models such as LDA do not need to know sentence boundaries, Tokup is able to reasonably robustly separate words without spaces between periods, while still keeping abbreviations in place in addition to expanding them and other acronyms based upon a large custom dictionary assembled from various Internet sources and creative efforts.

With word tokens in hand, the corpus was ready for LDA inference. Inference was performed using the Matlab Topic Modeling Toolbox implementation of *Gibbs Sampling*, with slight modifications to enable 64-bit processing (Steyvers & Griffiths, 2005). Multiple models were created by varying the model's parameters to produce a suitably understandable model. This level of subjectivity demonstrates the data modeler's role in creating a data portrait (Donath et al., 2010). We found that while increasing the number of topics allowed for more variability to be captured, it also became a tradeoff in overwhelming the users with too many potential options in the visualization itself. In the end we settled on 70 topics.

In choosing to represent the model on a 2D plane, there must be some reasoning on how the mountains get positioned relative to each other if they are not to be laid out a grid. We chose to position related topics near each other. To do so, we first calculated a similarity matrix between the topics using Kullback–Leibler divergence (KL divergence). KL divergence measures the similarity between two probability distributions, which in this case are the topic-word vectors φ . As KL divergence is not a true distance metric because it violates the required principal of symmetry, the KL divergences $KL(\varphi_1, \varphi_2)$ and $KL(\varphi_2, \varphi_1)$ are averaged. This similarity matrix is reduced from $K \times K$ to $2 \times K$ using MDS, providing the basis for the planar layout.

The visualization itself was implemented by scripting Maya, which allows for sophisticated rendering using soft shadows and easier camera movement. Post-processing text-layers were performed using Adobe After Effects. An interactive lower-resolution version was also created using Java using OpenGL bindings.

DISCUSSION

Here we can see the utility in having a common substrate or map that is shared across a variety of Twitter contexts. Often the instinct to summarize an individual's Tweets is to extract the relevant keywords and find a way to display them such as a word cloud. Here we find a common abstraction -- the LDA model of all of Twitter -- and use it as a basis to project any number of subjects onto it. With a static explorable map, observers should be able to carry over the cognitive skills that enable ordinary maps to be learned into a purely abstract visual domain. It is easy to imagine a widget that is embedded along the side of a Twitter user's profile such that we see their global topic distribution as a part of the user summary, which would be very helpful when seemingly random strangers begin to follow us (or in the context of Facebook, add us as friends).

The oddity of it is that the distinctions between topics and the global layout are in fact arbitrarily determined by MDS. Relative placements are of course meaningful, but there are no global insights that would tell us what going north or south would mean. MDS unnecessarily distorts the data portrait in this way by not using human-like methods of organization.

It can also be difficult to understand the topics themselves. The most probable words that belong to a topic might be readable by data scientists, and if coherent could produce a useful gestalt as is easy to do in Table 4.1, but the concept is very much esoteric to most observers. One option would be to involve human labeling for topics as the topic map itself is quite static. This approach is very appealing for models where the topics are very distinct and identifiable such as "health" or "science." However, Twitter captures a much wider variety of topics that may not seem Encyclopedic, such as the ordinary announcement of a user's current physical context (e.g. being at home, the airport, etc.). It also distinguishes in vocabulary usage surrounding a topic, from sentiment (good versus bad) to colloquial versus formal. Luckily, the 140 character limit in Twitter means that humans do much of the semantic compression, permitting models like LDA to find more useful correlations between co-present words.

The best part about LDA is that it requires no previous models of a given language. In this sense, it captures a lot of the nuance without bias from traditional formal corpus models such as those trained on newspapers like the *Penn Treebank* or stagnant samples such as the *Brown corpus*. Those corpora are not likely to cluster the words {*lol, haha, thats, funny, yeah, hahaha, ok*} or {*stuff, moving, office, box, place, pack, apartment*}, despite the common usage and interpretability of such sets.

Yet LDA is not without bias. It makes very strong assumptions in its generative model such as a) the order of words does not matter, b) words may exhibit polysemy but only little, c) all words are a member of a coherent topic, d) there are usefully limited number of topics that any one document "is about," e) variables such as time or authorship are irrelevant, f) all words may neatly fit into a small set of topics, g) most implementations use a symmetric prior on how often a word or topic is to occur. Clearly generous removal of stop words alleviates some of these issues,

but these assumptions remove a lot of the makeup that a human would use to socially and semantically distinguish documents and authors. Some models have been created to address some of these issues, such as LDA-HMM (Griffiths et al., 2005), Author-Topic model (Rosen-Zvi et al., 2004), and Pachinko Allocation Model (PAM) (Li & McCallum, 2006), but those issues are outweighed by the other issues if a true human-like thematic capturing is desired.

LDA also cannot capture meta-level themes that go beyond domain. For example, sarcasm or humor would be very difficult to encapsulate in a topic model unless the specific form of tone unusually uses highly specific symbolic words. Similarly qualities such as arrogance would so be nearly impossible to capture. Further, the delegation of topic may miss important distinctions in the domain-specific choice of words. For example, LDA may very well cluster web browsers together, but those who talk about Microsoft's Internet Explorer would be judged by the technical vanguard negatively compared to Firefox or Chrome. Identifying the topic alone is not enough without further qualification. Those judgmental qualifications are difficult for any language model to capture simply because the "so what?" judgement or perception of reading any one tweet is not available.

Finally Landscape of Words is unable to show coherency within a given topic. Similar to the web browser example above, the visualization lacks the ability to communicate how a topic becomes split across the tweets that underlie it or the correlation between topics beyond spatial proximity. Are there two main factions across the tweets (e.g. *Red Sox* versus *Yankees*), or does the word choice matter much less (e.g. *haha* versus *hehe*)? Hierarchical topic models such as PAM may help with such a task, but tests have revealed that forcing symmetric hierarchical splits yields poor results. Future models should find the ability to split more organically the topic into subgroups, independent of the number of splits for sibling parent topics.

SUMMARY

Landscape of Words tries to expose the underlying themes within a community, and uses them as a backdrop for portraits about each user and their networks. Its fundamental tool is topic modeling, which is able to extract socially meaningful textual clusters from a corpus of millions of Tweets. These topics are then visualized for observers using a topological map metaphor. Topics are identified by their most probable words, and related topics are placed near each other. The map alone reveals the trends of the community at large, but when combined with heatmaps also can create data portraits of individuals and collectives by projecting them onto the common topic space. While the interface may lack some practicality, Landscape of Words demonstrates the potential to prototype subjects within larger cultural trends and sociolinguistic features.

4.3 Experiment #3: Personas

In our current and future world, digital histories are as important if not more important than oral histories. It is not just the digerati who leave behind vast footprints of identity in various fragments and form; children are beginning to construct their (social) identities through archived media at young ages (boyd, 2007). Already a reality in advertising, health care, and terrorism intelligence, fortunes are sought through data-mining vast information repositories, making the computer our indispensable but far from infallible assistant. *Personas* is a data portrait of statements about a subject's name on the web and their machine-generated characterizations. It demonstrates the computer's uncanny insights and its inadvertent errors, such as the mischaracterizations caused by the inability to separate data from multiple owners of the same name. It is meant for the observer to reflect on computational methods of condensing our digital traces given they are opaque and often socially ignorant. By making its data processing transparent when it is normally opaque, *Personas* exposes the inhuman side to machine learning. At the same time, it also provides a method to aggregate heterogeneous textual information about a name into a single graphic. *Personas* was designed with the input of Greg Elliott and the Sociable Media Group (SMG): Alex Dragulescu, Yannick Assogba, and Drew Harry.

PROBLEM

In 2009, the MIT Museum solicited an exhibit from SMG. After much planning, we created Metropath(ologies): a show about living in a world overflowing with information. Much of the exhibit centered around issues of identity, privacy, and surveillance, guided through optimistic and dystopian lenses. Largely an installation of tall projected columns, museum-goers would be surrounded by flashing bursts of images of people earlier in the show, simultaneously in the show, or across a portal to a virtual city (Figure 4.11). These cities were inhabited by varying aspects of FriendFeed members. With the rest of the pieces focusing on others' histories, *Personas* was created out of the need to pull in the museum-goers own digital identity (see Figure 4.12).

Personas seeks to question what it means to have personally-identifying digital footprints across the web. While many online activities take place under the mask of a handle, there are many places that do mention real life names (often without informing those who are mentioned): newspapers, corporate directories, soccer leagues, speaker biographies, home pages, sexual offender lists, personal (but public) blogs, Facebook and Twitter profiles, and more. It is common to Google someone before going out on a date, just as to rely on a service to find flaws in social networking profiles of potential employees (Bell, 2011). With a future of increased sharing (Hansell, 2008), *Personas* provokes individuals to consider their existing presence and the perspective of a machine automatically classifying them in a non-human form.



Figure 4.11. Metropath(ologies) columns and data surround data artist Alex Dragulescu. Images from the news, web, and gallery surround museum-goers as they traverse the columns. On the periphery of the columns laid the other data portraits including Personas.

Search Engines

Search engines are currently the predominant way to find public information about a given name. While not the only way, search engines are worthwhile to critique given their dominating paradigm in practice and thus in thought (McLuhan & Fiore, 1967). The structure of search engine results are such that a given name is mixed in a variety of results: social network profiles, statements of character and biography, sports scores, address records, etc. Search engines do not visually split apart the types of information found or summarize them as some people-focused search engines try to do (see Figure 4.13). They also do not distinguish between those with a shared name.

Instead of a person-centric representation, impressions about people are generally created from whatever happens to be within the first few pages of a Google result. We have trust in Google to bring forward the most “relevant” results, as it does for most of our other queries. However,



Figure 4.12. Museum-goers using Personas at the MIT Museum.

information about individuals are likely to be scattered across a variety of possibly low-ranking sites such as the local high school newspaper. Without regular effort to dive deep into the set of results with mixed qualities and names, we are likely unaware of the potentially vast information available online. Such information is very visible to those seeking deep impressions including services like Pipl or those who might wish to make a very informed decision, possibly nefariously.

Individuals reading through returned search results process the available information much differently than a machine. Yet when machines seem to work well enough, we assign too much authority to their results. The inner methodologies in machine learning, while informed by humans, most certainly do not reflect actual human-level reasoning. The working paradigm is to build simple algorithms that encompass human behavior for a given interface, as opposed to virtually replicating a human-mind and its operations (Halevy, Norvig & Pereira, 2009). A human is likely to come to a very different conclusion as to how to surface available information compared to Google. Yet because the results are seemingly good enough, we project more humanity and trust onto the machine than we should (Reeves & Nass, 1996).


Search engine results are bottom-up lists of data rather than top-down summaries. As such they skirt around our expectation for a germane human summary of another person. Because they do not need to make sense of the returned data, they avoid our potentially harsher judgement of

pipl Name [Email](#) [Username](#) [Phone](#) BETA [Business](#)


[Clear](#)
First Name Last Name City State Country

Aaron Zinman, Cambridge, Massachusetts, United States


Background Reports

 **Aaron Zinman, Cambridge, MA... \$\$\$**
 Public Records & Background Checks - Intelius - Sponsored Result www.intelius.com - Deep Web
 Sponsored Tip: Other sites charge \$39.99 for just one report. Run unlimited checks free... [BeenVerified.com](#)

Personal Profiles


 **Aaron. 30, CAMBRIDGE, MASSACHUSETTS...**
 Personal Web Space - MySpace www.myspace.com - Deep Web
 Sponsored Tip: Find hidden profiles and photos across MySpace, Facebook and 40+ networks... www.spokeo.com

Email Address


 **Aaron Zinman [A*****N@...COM] , Cambridge, MA... \$\$\$**
 Email Address Records - Intelius - Sponsored Result www.intelius.com - Deep Web
 Sponsored Tip: Find out Everything About Aaron Zinman in Seconds... [BeenVerified.com](#)

Results for Aaron Zinman, Massachusetts, United States without Cambridge

Professional & Business


 **Aaron Zinman, MA, US, PhD Student at the MIT Media Lab, ...**
 Professional Profile - LinkedIn www.linkedin.com - Deep Web
 Sponsored Tip: See Who's Searching for You at America's #1 People Search site... [MyLife.com](#)


Publications


 **RadioActive : enabling large-scale asynchronous audio discussions on mobile...**
 Scientific Publication - Scirus www.scirus.com
 Sponsored Tip: See Who's Searching for You at America's #1 People Search site... [MyLife.com](#)


Results for Aaron Zinman without United States

Personal Profiles


 **Aaron Zinman...**
 Personal Web Profile - Facebook www.facebook.com


 **Aaron Zinman...**
 Personal Web Profile - Facebook www.facebook.com


 **Aaron Zinman...**
 Micro Blog - Twitter twitter.com - Deep Web


 **Aaron R Zinman...**
 Customer Profile - Amazon.com www.amazon.com - Deep Web
 Sponsored Tip: Access private profiles and photos on MySpace, Facebook and 40+ networks... www.spokeo.com

Videos


 **zjemily persona. 1 min** Aug 19, 2009 Uploaded by zjemily, And my user's metropath, ...
 Videos - YouTube www.youtube.com - Deep Web


 **Zu ND Odyofilz persona. 33 sec** Aug 19, 2009 Uploaded by zjemily, neat programming...
 Videos - YouTube www.youtube.com - Deep Web


 **characterizing mark guadalupe ; -).** 59 sec Aug 26, 2009 Uploaded by mak1e...
 Videos - YouTube www.youtube.com - Deep Web


 **documenting a (web) persona. 2 min** Dec 4, 2009 Uploaded by nicomachus1, MIT labs and...
 Videos - YouTube www.youtube.com - Deep Web
 Sponsored Tip: See Who's Searching for You at America's #1 People Search site... [MyLife.com](#)
[3 additional Videos »](#)

Publications

 **[PDF] RadioActive: enabling mobile-based audio forums [PDF] from psu.eduA...**
 Publication - Google Scholar citeseerx.ist.psu.edu

 **Navigating persistent audio. A Zinman* - CHI'06 extended abstracts on Human factors...**
 Publication - Google Scholar portal.acm.org

 **Data portraits [PDF] from mit.edu, J Donath, A Dragulescu, A Zinman, F Viégas* -...**
 Publication - Google Scholar www.mitpressjournals.org

 **[PDF] Is britney spears spam [PDF] from psu.edu, A Zinman* - Fourth Conference on...**
 Publication - Google Scholar citeseerx.ist.psu.edu
 Sponsored Tip: Is Aaron Zinman looking for you too?... [MyLife.com](#)

Sponsored Links

Keep your private data private with MyPrivacy Reputation.com

Background Check Them Instant Report Free Trial [BeenVerified.com](#)

Over 50,000 people find out every day [MyLife.com](#)

See Private Profiles on MySpace, Facebook + More [Spokeo.com](#)

Aaron Zinman. See User Ratings of Record Sites. [PeopleSearching.com](#)

\$1.95 Phone, Address And more. Preview Results [USSearch.com](#)

Figure 4.13. Commercial person-oriented search engine Pipl.

its top-down capacities. It is easy to assume only the relevant links have been listed as we see no evidence of non-intelligent behavior. Those links left out of view are rarely seen enough to know otherwise.

Human Modeling and Presentation

Instead of an easily interpretable, multi-dimensional, interactive, and nuance presentation of a set of data, the standard catholic paradigm of bottom-up listing is the easy solution when an internal representation is too esoteric to be legible. Just collapse everything into a single dimension, so says the mentality. The hope is that the simplicity of a one-dimensional interface outweighs the complexity of a more nuanced perspective, and that any error in this drastic oversimplification is tolerable. It is the fault of the secretive and arbitrary internal models that we are left with an inhuman approach to data. The conversation is asymmetric: the users are subject to the capacities of the interface designers and data modelers. The reason is two-fold: 1) the secret black box is seemingly worth more as intellectual property than a more exposed model, and 2) machine learning uses models that may be too difficult, mathematically-inclined, and at the wrong level of semantic granularity to easily communicate them to users.

While the data modelers may have good reasons, the user is left to fight aligning the given representation with against their own interests and goals. The hows and whys of an informational interface are nonexistent, giving only the faint implicit answers. We do not know why Netflix makes “Strong female lead” a category but “Strong male lead” is not. Consequently, we cannot ask Netflix for their strong male lead recommendations. The users have little say in reorienting a model and its presentation to suit their goals. In reality, the presentation and model is often more constructed in a particular way because of what was easy or popped out of the data using standardized techniques, rather than an explicit decision to the most human-like representation.

As a result we take the existing presentation and construct our realities around it, projecting onto it new capabilities and insights that are in reality ill afforded. The desire for a low-cost impression battles against what is possible using machine learning and non-agent-like representations. The ignorance in the common populace of how these systems are built and what value can be extracted only further skews the perception and thus usage of technology. Because hidden data remains offscreen, we incorrectly understand how balanced an impression or representation might be in comparison. In turn, our false assessment of objectivity and authority becomes inculturated to the point where machines are given far more benefit of the doubt than they deserve. The danger comes when we trust computational models to make critical decisions at scale using reasoning processes that are flawed in their assumptions.

Control

Those who write algorithms that compute individuals often do so without the explicit permission of people being computed. Without knowledge or participation, they do not have say in how they are being internally represented, what should be ignored as too private, what is germane, and any meta-estimate of the error due to sparsity and other related issues. Complicating these factors is that those with large digital footprints give the illusion to the data modelers that an apparent density provides a balanced perspective as they create their models. This is rarely the case: too many important aspects of an individual will be missing are too difficult or non-obvious to analyze.

Regardless of sparsity, the classifiers that determine who is a spammer, who is an influencer, who is a terrorist, and who is a “good” employee are all the result of a set of arbitrary choices that could just have easily come out somewhat different if another algorithm or feature were used. For example, Klout currently assigns “*viral value*” to social network profiles based upon a number of arbitrary factors they’ve determined are worthwhile. If they included IQ, best linguistic practices, FICA score, and a curated score by Oprah Winfrey their final scores would surely be impacted. Yet their decisions until these factors are added will become an anchor point regardless of sophistication. These axiomatic choices collectively define the current world. There is a large risk when society as a whole is unaware of the power that is being used on them, possibly using models dangerously when decision makers do not understand the limits of the intelligence that underlie their reasoning.

DESIGN

Personas exists to expose the inter workings of machine learning in an entertaining form. It mimics the larger backend practices by visualizing a modern and representative algorithm in its characterization of an individual. It does so by scouring the Internet to find any information available about a given name. Enter the name, and outcomes a scurry of colored lines shifting as the machine applies its stochastic inference process on what was found. In a public artistic context, Personas becomes a digital portrait of any publicly accessible name.

It was originally designed to be a standalone piece in a public museum (it later became popular when put on the web). While any museum-goer could choose a name, it becomes a public spectacle whereby the nearby audience collectively joined the machine in judging an individual by the surfaced information. The desire was to strike a balance between a pleasing neutral display and a very authoritative yet subjective partitioning of the individual in abstraction.

The first screen of Personas, as shown in Figure 4.14, asks for a first and last name. A brief description underlies the web-version, whereas the museum-version only simply asks for the

name. An automatic countdown allows the user to progress if they do not press enter. *Personas* was designed to be driven solely by a keyboard.

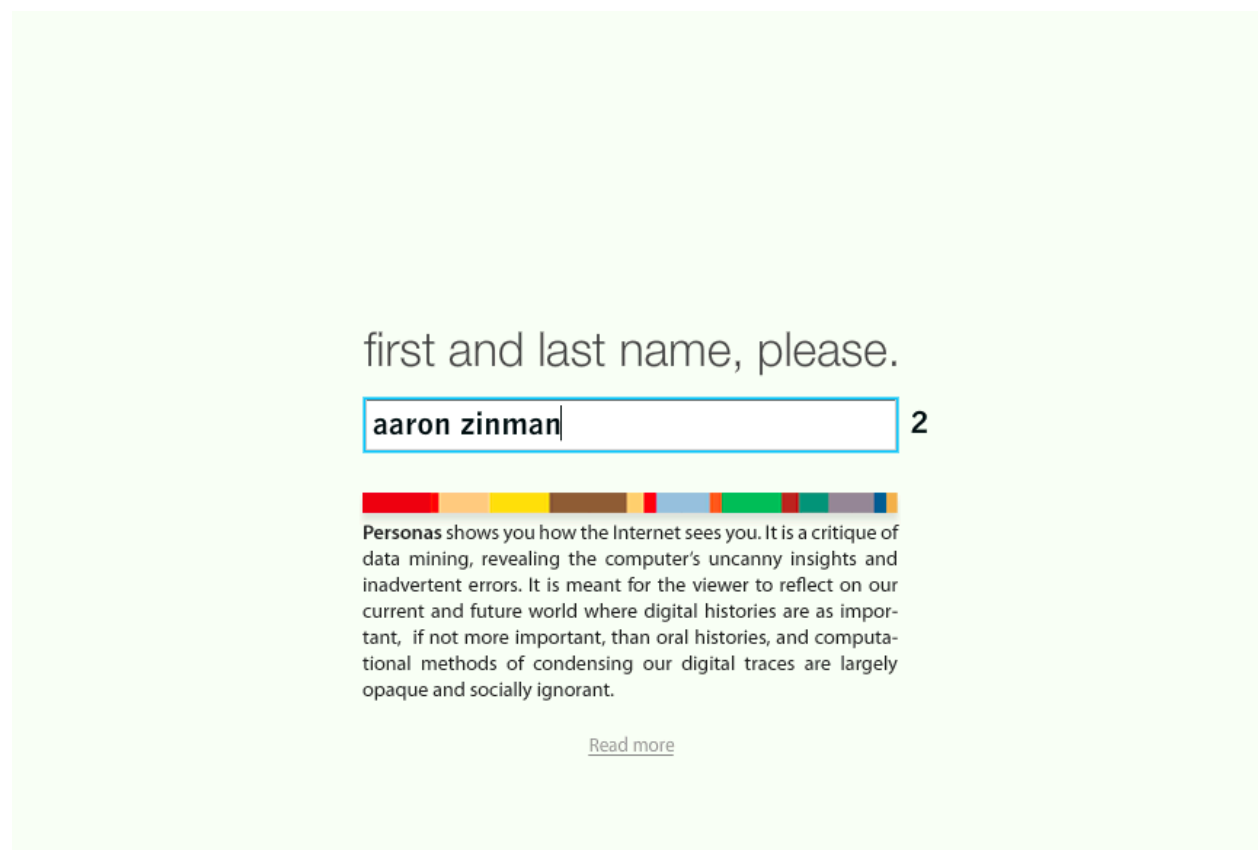


Figure 4.14. The entry screen. Individuals are invited to enter a full name, which could be their own or another individual.

Soon the interface says *“finding out what we can”* while the backend searches for the name, performs the characterization, and serializes it to the client (see Figure 4.14). Quickly the screen dims and flashes to the text retrieved as it visually performs the first iteration of analysis. The text comes from querying the larger body of the web in a decontextualized manner using a linguistic hack. Each name that is entered is in turn searched for statements about them in the form of *“first name last name <to be conjugation>,”* for example *“Aaron Zinman is”* and *“Aaron Zinman will be.”* The extracted data is context-free; only complete sentences are cherry-picked from their housing webpage without regards to the surrounding text. This is not so far off from many typical machine learning approaches or the process of Googling someone’s name: we see lists of results in a similar manner. The returned text is characterized in front of the user by visualizing an algorithm called Latent Dirichlet Allocation (LDA).

Personas was designed to visualize LDA using a Gibbs sampling inference process. This means that the underlying algorithm is stochastic -- it uses probability to make its decisions in a way that

two identical runs may have different results -- as well as iterative due to its use of Monte Carlo Markov Chains (MCMC). Thus the visualization has the possibility of not showing just the final conclusion, but an underlying process to arrive at a conclusion. This is the basis that allows Personas to be an experience seeing the machine reason, rather than the typical input:output infographic that other algorithms afford. LDA is used to infer new statements against a predefined model, consisting (to the user) of high-level categories such as travel or management.

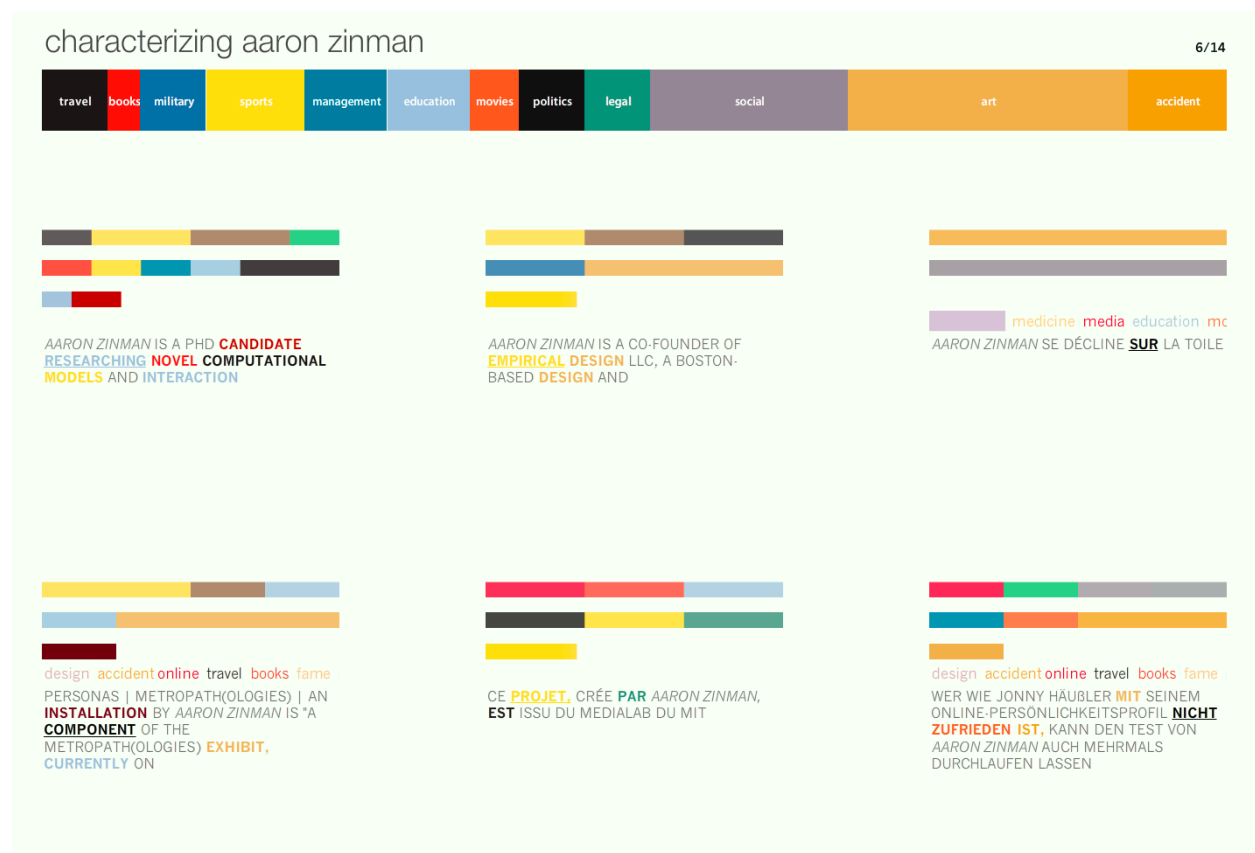


Figure 4.15. The main experience of Personas. Each known word is cycled per iteration judging its category memberships, resulting in its final color representing assignment to a unique category.

Thus Personas is visualizing the characterization of individual words and thus statements as a whole to a predefined model, resulting in aggregate a weighted vector representation of an individual based upon the available data. Collapsing individuals into a pre-defined model, regardless of model suitability, is the standard methodology for machine learning in practice. While the better data modelers attempt to make their system as catholic to the data as possible, this is a difficult task. Any error then present in Personas is consciously part of the design; the individual probabilities demonstrating confidence are purposely not shown. The viewer is meant to reflect on its errors as well as its successes.

The underlying visual logic is simple and mostly a 1:1 mapping with the inference process. Each word familiar to Personas (skipping stop words, etc) is underlined as it attempts to match it to a given category. The names of the categories fly by until it makes its selection to represent choice by the machine, finally coloring the word to match the chosen category in the ontology revealed above. A colored strip lies above the sentence being analyzed. They are created in synchrony to represent the analysis of the individual words, as can be seen by the not quite complete bottom strips in Figure 4.15. By constraining the total strip width to be the same for each data point, the width of any one bar is proportion to the number of known words in that sentence. When two words are characterized into the same topic, those words are then grouped together in the visualization to create a longer single color block. The conclusion of each sentences iteration are timed to end together, resulting in the update of the master strip at the top. The previous iterations are retained on the screen so that any oscillations in machine thought are displayed, showing the ebb and flow as the iterative stochastic algorithm reaches its final conclusion.

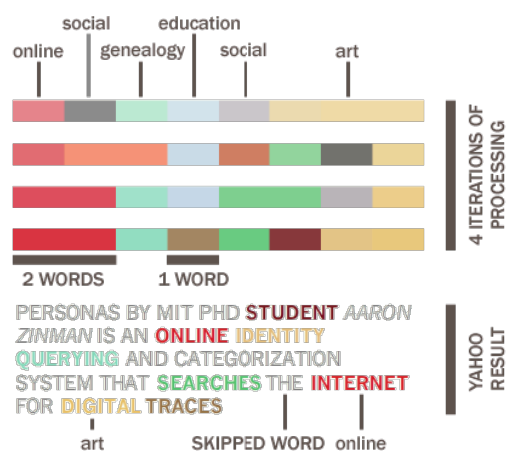


Figure 4.16. Visualizing the analysis of a characterizing statement. Each colored block represents the mapping of a known word to a pre-defined category. Longer blocks represent multiple words. Because the underlying algorithm is iterative, previous mappings are retained above the final conclusion.

Personas continues to analyze possibly more statements than visible on the first page by sliding in the next set. A counter at the top right depicts how many results have been processed out of the total. After all sentences have been characterized, the final set fades out as the master strip slides down to reveal the final weighted vector representation of the individual according to its predefined model.

Personas uses the visual language of statistics to question the authoritative presentation dominant in most interfaces today. Here the user cannot select which sentence belongs to them and which to someone of the same name. They are all clustered together into the same model. The problem of canonicalization is hard, as many know who the TSA wrongfully flags, and is a consequence of a data model-driven world until everyone's action are always tagged by their unique

DNA. As the results stream by, the user is not given a chance to annotate the results, recognize temporal relevancy, or to disagree with the model characterizations to the words. The resulting strip then becomes a data portrait to which one can debate. We cannot debate what is seen in a true mirror, but we can easily debate abstraction of identity.

The danger lies in the haphazard projection of identity onto a representation that does not deserve it. Should the model code a data set incorrectly from an objective sentence-granular

point of view, the final vector may still accurately reflect a higher-level self-identity. The opposite may occur as well, because in putting *sports*, *management*, *music* next to a name, we detach ourselves from the limited sparse corpus in subject and into more seemingly complete prototypes. Personas purposely provokes the user by ignoring the issue of sparsity in its conclusions.

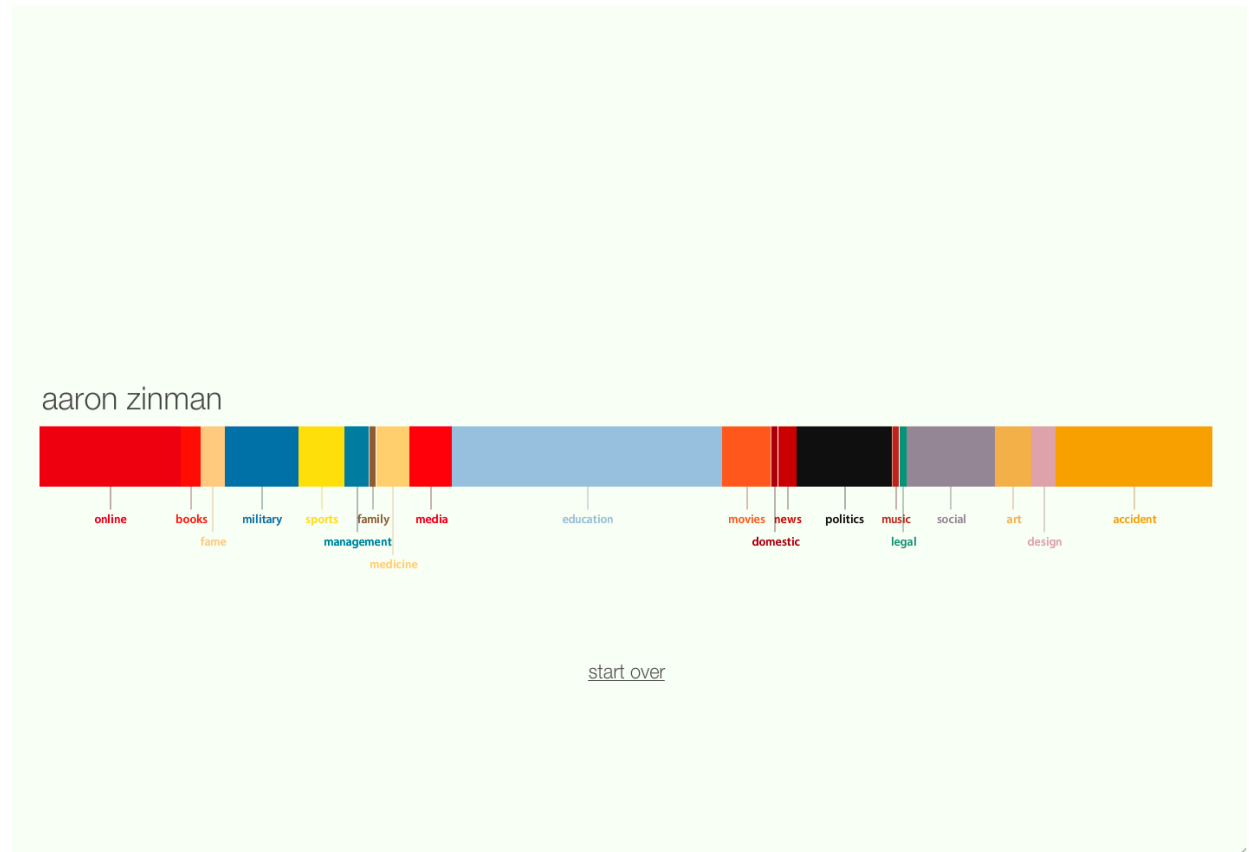


Figure 4.17. Final screen of Personas: the resulting characterization of and individual. We see the final weighted choices made by the algorithm in decomposing the found information into a set of predefined categories.

IMPLEMENTATION

Personas can be decomposed into 3 main systems: 1) name factoid search, 2) factoid characterizations, 3) the front-end to visualize the results. The search and characterizations are implemented in Python. The front-end talks over HTTP to the backend using a REST API powered by CherryPy/WSGI/Apache at first, and then later replaced by Facebook's Tornado web server. The front-end was implemented using Adobe Flash/Flex, allowing both web-page embedability as well as AIR's implementation for a museum. A custom OS X kernel extension was implemented to disable control keys including escape to achieve a full Kiosk mode. We now break down each system.

Name factoid search

Given the task was to find available information about a name, using a search engine naturally became a logical choice. We settled on Yahoo's BOSS due its quality of service, free cost (at the time), and unlimited query rates. However, searching for a name alone presents a challenge: many times names simply occur in lists or are otherwise detached from a meaningful context. Because the goal was to find characterizations, we employed the language hack of "*name <to be conjugation>*" as previously described. The resulting text by itself were found to be generally interesting statements, assuring a solid foundation on which to build the rest of the analysis. Figure 4.18 demonstrates sample query results. Once the data is retrieved, it is post-processed to extract sentences using custom regex-powered heuristics, then checked for duplicates, violations of the exact search rule, links to name farms or social networking profiles, general garbage results, non-English results, and non-complete sentences. The finalized sentence is saved in its raw form, then further prepared for NLP by aggressively removing stop words and using the Porter stemmer.

John Woo

John Woo is generally regarded as the first Asian director to find a mainstream commercial base
John Woo is a musician best known for his work with the indie pop band
John Woo is making movies written by the Wachowski's
John Woo is one of my absolute favorite songs, and that's just my personal take on it.
John Woo will be in Singapore to talk about his new film Reign of Assassins on October 1st.
John Woo was cool, or, I like subtitles
John Woo was born in China and has an estimated net worth of \$60 million dollars
John Woo was not making a movie to for the plot/character development

Caroline Smith

Caroline Smith is a singer-songwriter from Minneapolis, Minnesota who started recording and performing at age 15, opening for B.B. King
Caroline Smith is hungry for your confessions
Caroline Smith was not far behind as she twice smashed her personal best
Today is the day when Audrey Caroline Smith will be delivered into this world and then...barring miraculous intervention...into her eternal home
Caroline Smith will be traveling with her band
Caroline Smith is Fucking Insane
Caroline Smith is originally from Detroit Lakes, Minnesota, a town of about 8600 according to the 2010 Census

Figure 4.18. Results of Personas-style Yahoo search queries on two random names, John Woo and Caroline Smith.

Factoid characterizations

To characterize the text, we decided it would be best to organize it according to natural rhythms found name searches. Thus the data is not being measured to anything but what is found in by the millions already online. In order to characterize an individual, we needed to gather a large sampling of the baselines for comparison. To do so, we simulated individuals using Personas by generating two million random names. US Census data on the frequency of last names was used to generate realistic American names. Those were paired with randomly selected first names compiled using various Internet resources. The generated names were searched in the same manner as Personas normally queries for users (each *to be* conjugation as a separate exact matching search) and gathered the results into a central repository. The resulting corpus of over 10 million factoids was ready for analysis.

In looking for natural rhythms, it is best to use unsupervised clustering methods to let the data speak for itself. Given the textual data, LDA was a natural candidate. As the task was to gather information on an individual, and organize it to produce a characterization of them, the requirements well fit the notion of topics being a latent representation of an entire corpus down to individual or collections of documents.

The corpus was further paired down by aggressively removing stop words, names, too short documents, and more into a cleaned 2.7 million document set. Those documents were stemmed using the Porter stemmer, and finally clustered using Mallet's implementation of LDA to produce very legible results (McCallum, 2002). The seven most probable stemmed words are shown next to the final human-generated title for each category in Figure 4.19.

When a user is searching for a given name, the name is searched and post-processed as described before. To prevent Personas from getting too repetitive, large result sets are capped at thirty by random selection. Category membership is finally inferred using Gibbs sampling of the common LDA model, generating and sending intermediary and final characterizations to the client serialized using Google's Protobufs.

Front-end

The frontend was written in Adobe's Flex environment. It communicated with the backend over HTTP requests, receiving a Protobuf serialized message describing the raw results and each stage of inference on each sentence. Personas makes heavy use of the TweenMax library to perform its animations. It was found that using opaque colored DisplayObjects with bitmap caching achieved the greatest performance when compared to other methods of compositing and drawing, such as relying on the Sprite class or direct bitmap operations on a single drawable-object. In a museum-context, a client-generating heartbeat and python-backend receiver were used to kill and recreate the AIR process should an error occur. In combination with the custom

kernel extension to prevent Escape from exiting full screen, the self-monitors ensured reliable operations in public settings.

Figure 4.19. The seven most probable stemmed words per collapsed named topic in Personnas.

DISCUSSION

Personnas instantly connected with audiences. At the MIT Museum, it gathered large crowds as shown in Figure 4.12, where half the experience was personal and egoistic, and the other was to witness public curiosity about someone else in an amusing social context. What will it say about *me*? As Personnas was implemented in Flash, it was easy to put online. After being picked up by a blog called Infosthetics, it quickly spread, so much that it has over one million uses in the first month and well over two million as of this writing (and counting). It has long stabilized at between 500-1000 hits per day. Needless to say, Personnas seemed to have struck a chord.

Why was Personnas so successful? After speaking with many individuals and witnessing its usage, it seems to stem from three main points, 1) narcissism, 2) consumability, and 3) shiny factor.

Humans are naturally curious about themselves, which also happens to be their favorite conversation topic (Carnegie, 1936). Will it find the real me? Will it be positive? What conclusions will it come to? At the end we are given the chance to judge the defenseless machine in its errors and insights, another favorite human activity. Because the whole experience is less than four minutes, it has more of a bite-sized entertainment factor. The package could be replayed for those more curious as to what is happening, whereas the less impressed feel they at least achieved some conclusion.

It helps that the design of Personnas was well received. Its hypnotic slot machine-esque animations and use of color gave novelty to a normally static result-oriented web. Audiences also seemed apt at reading the final representation, understanding the language of statistics to know for example that a bar with the name illegal can be bigger than one might like. Other parts such as the sub-iterations and text coloring did not need to be understood as deeply, and for many successfully became part of the background gestalt.

Personnas lacks obvious utility. It would be difficult to take the final characterization and find value in it alone. The returned snippets can be also pulled from a diligent Google search. In its museum-form there is no explanation of its intended purpose or process. When put into the web detached from an art museum, its utility is even more questioned. However utility is not the point: it is a provocation about digital identities, about its sophistication and flaws. It is a data portrait that is as much a caricature of the data modeler as it is the subject.

Here the seemingly arbitrary choices of machine learning color the possible outcomes in representation. The predetermined number of topics could have been increased or decreased by

online

comment email view add repli regist mail
video photo blog news view imag share
blog search comment rss content skip question
voic het een met die month voor
receiv compani network connect world busi commun
web site contact design inform home search
forum messag member view discuss search board

travel

travel hotel locat island road time inform

books

book review custom help order search ship
book author fiction novel romanc review mysteri
book author publish writer award children photograph

fame

photo news famou born celebr pictur gossip
movi review dvd film video game onlin

military

war army forc serv air unit command

sports

game team coach play player year point
poker award year world olymp state game
game goal play match team minut win
game team nfl yard defens year football
basebal game leagu hit team pitcher pitch
club news football golf team player leagu
athlet site offici univers network state coverag
race year seri time car team event
team coach year state senior women school
football sport colleg basketball coach player news

management

research director develop program commun project educ
manag presid busi director develop compani consult
director chief presid execut offic appoint manag

genealogy

born marri famili died genealog daughter tree
born counti famili marri father live daughter
church born servic funer home mrs memori

fashion

design food wine shop product home garden

family

born died marri counti children apr child
year life time book cancer famili live
year born mother famili father daughter baby

committees

meet board member committe presid minut chair

aggression

fight wear year night time red got

medicine

inform problem health test articl need weight

media

radio event host ticket news broadcast talk

domestic

city build area year hous counti locat

education

train class year school program student learn
univers student colleg award school year graduat
univers book histori professor american studi author
school year student teacher graduat univers educ
univers professor research engin scienc studi depart

movies

imdb credit report manag error alongsid databas
imdb movi news credit celeb alongsid messag
film actor play produc director movi televis

news

news local sport weather counti sourc inform
articl editor news magazin issu publish archiv
news sport local busi latest entertain break
news newspaper onlin journal press san commun

illegal

charg court year arrest counti state convict
polic offic year counti arrest sheriff kill
year murder death kill charg polic prison

politics

polit right report world american news peopl
elect vote candid counti city parti mayor
state obama democrat senat presid elect unit

music

music band record album artist play jazz
music ebook shop download compar review album

legal

licens year state rate review phone inform
legal attorney lawyer state firm case practic
court counti state city district attorney case

social

group member meetup meet sinc introduct titl
year life time live thing girl peopl
realli time thing hes ive peopl got
time year night home morn got hous

religious

church pastor ministri year rev servic speaker

medical

health care medic hospit center nurs servic

religion

life author book peopl world spiritu speaker

professional

busi market compani manag servic year financi
real estat home agent list realtor search

musical

music perform danc compos concert theatr jazz

art

artist design paint galleri photograph exhibit arts

design

score rate submit averag member year shirt

accident

kill accid crash car und die der

six without much change to the overall aesthetic⁹. Similarly, the actual human-given name could be altered in emotional charge without semantic shift. For example, the current category *illegal* could have just as easily been named *court* or *legal* for the given top probable words.

The choice of wording unfairly biases the most probable words over the remainder. As the raw probabilities are never revealed, the viewer is unaware of how well fitting their unseen text is to the model. Some topics tend to have some semantic category at its top probability percentiles, but a different set in its 80% percentile. Given the semantic shift, the category name assignment is no longer appropriate and improperly biases the observer. This particularly occurs in LDA when the model is short of the number of “true” topics.

Personas could have revealed confidence values in a more legible form, but the mystery behind it emphasizes its authoritative stance despite errors. If it exposed more of its underlying model, observers may not understand the underlying divergence, only confusing them. Keeping a single high-level word eases its interpretation given there are 31 categories in total. If we showed the top five words per category that would total 155 words to interpret in addition to the raw text. Here abstraction wins out in representation for the user. A more utilitarian piece could let the user drill down into the underlying data.

Watching people use Personas, I am not convinced many would drill down; people tend to take in Personas in a single distorted impression without much further considering. They tend to divorce the data from the characterizations strip, where they read the data but then judge the strip independently. Despite the raw data being present, audiences tend to project themselves into the characterizations, and in doing so, assume much more of the capacities of the algorithm than deserved. It was not uncommon to hear “*Oh, 15% sports? Well I did play sports in 4th grade so that must be where that is coming from.*” The lack of any sports-related sentences seemed to be unrelated. The observer will tend to see what they want to see, especially as the categories and weighting are vague and non-referential. Because the final form was not 1:1 with the raw data, abstraction removed the understanding of the link and thus allowed for a different impression of the conclusion to occur.

When a common name occurs, individuals often reply to the screen “*this is me*” and “*this is not me.*” However, Personas does not allow for such corrections, or any corrections much less a deeper dive. If this had been allowed, would that have brought satisfaction to the users? Would they see it as more trustworthy? It is not clear given the final representation still may not match their own projected identity in the end. Some nuance or key component will not be captured, whether in the data found, in the weighting of the individual pieces, or in the projected model. Giving user

⁹ In this data model, very frequent topics such as sports were merged by hand. While setting the number of topics greatly affects the outcome in LDA, symmetric priors pressure the model to let popular topics take the place of less frequent topics.

feedback only provides an illusion of control or freedom; ultimately any control still remains within the constraints of the system. Once users start to play around, they will very quickly realize the extent of capabilities once its limitations are quickly reached. It is not a human, and thus you tell it to make exceptions or pull in data from outside biases. Yet the raw text is uniquely human, providing total mismatch between the inputs and outputs.

While it is obviously possible to build a representation far more sophisticated than Personas, the piece contends that trying to define the human in an interface may be impossible. There will always be some aspect missing. A dimension that is much more important than the others. Despite the tendency to anthropomorphize computer interfaces (Reeves & Nass, 1996), the more users are able to push the synthetic boundaries the more the separation of intellect will become apparent. Personas attempts to visualize one algorithm that can be easier to understand than many others. Yet it fails to convey to all audiences the underlying mechanics; how far can we reveal the underlying processes and still be usable and legible?

The future of social data is being currently fought by the power players in Silicon Valley: the Googles, the Doubleclicks, and the VC-backed Facebook monetizers. With an unawareness in the public about the life of data and its potential future value, more systems like Personas are needed to expose the risks and opportunities along with a balanced perspective on bias and representation. In challenging the illusion of ephemerality of personal data and the nuances of representing individuals, Personas provokes its user to contemplate their effects of their own past actions on the future reputation.

REACTIONS

As previous stated, once Personas was written about in information visualization blog Infosthetics, its popularity on the web skyrocketed. At its peak Personas received 69,311 hits in one day. As of this writing it has been visited 2,382,015 times, and currently averages 400-800 hits per day. Over its lifespan as well as in the past three month, 82% of visitors are new while only 18% are return users. They spend 1.56 minutes on the site on average, although this number is skewed because 60% of viewers leave before 10 seconds. The majority of actual users spend between 3-10 minutes on the site. There are thousands of people who have used Personas over 200 separate times, although they make up less than 1% of the total visitors. Most users find out about Personas from a webpage or blog, making up 65% of the traffic. Approximately 31% uses direct links, which most likely come in the form of users emailing to each other. Most web searches come in some variant of *personas* and *mit* or *media*, suggesting a possible institutional bias in interpretation of the work.

Many, many blogs and newspapers have written about Personas including CNN, PBS, The Washington Post, New Scientist, The Guardian, MIT Tech Review, TechCrunch, ZDNet,

UTNE Reader, Coolhunting, IBM Developer Works & Research blogs, and more. It also experienced a true portrait-like experience: many users started uploading screenshots and videos of their name to Flickr and YouTube. Many of the comments by users were similar, so we performed genre analysis across a sampling of top referring traffic, low-referring traffic, personal blogs, and high traffic sites, with 17 in total. The sites description and comment on Personas, along with its comments were codified using the surfaced genres. Table 4.3 shows the summary of genre incidents across the 17 sites.

Total Incidents	Genre
26	Fun / cool /interesting
10	Name canonicalization
8	Inaccurate
8	Different results each time
8	Confused
5	Anti-climatic
6	Accurate
5	Not practical
5	Disconnect between text & category
3	Scary / worried / see the warning
2	Less data than google
2	Reiterates concept
1	Talked about as tool
1	Different amounts of data per pseudonym
1	It breaks out of black box paradigm
1	Will reuse portrait in other contexts
1	Results are other people
1	Surprised
1	Learned something new in the data

Table 4.3. Total number of incidents of a given genre found in the various comments about Personas in blogs, both major and minor.

The main sentiment across the comments cast Personas in a favorable light, calling it “*really interesting*,” “*beautiful design*,” or just plain “*cool*.” The next biggest issue was simply that of name canonicalization is not performed in Personas, so many users found data not relevant to themselves. While this is part of the intended experience, many were put off nevertheless. People often complained of its stochastic nature, expecting some stability, and of its inaccuracies. A few reflected about the meaning of Personas in a non-utilitarian sense, although most seemed to expect it was a tool.

Perhaps much of the confusion about Personas was due to its divorced context from the museum. However, in doing so the reactions test the water for future researcher who desire to make Personas into a real tool. Appealed by the digital mirror, users wanted a context to evaluate the final strip against. An actionable output would provide reward for having watched the process. They also wanted to be able to manipulate it to identify which were them, and correct errors made in categorization. Finally, while many understood the weighting of the bars, they did not always understand how they came to be. More explanation needs to be put in place for users to know what a data mining engine looks like.

SUMMARY

Personas is a component of the Metropath(ologies) exhibit, originally on display at the MIT Museum. It uses natural language processing and the Internet to create a data portrait of subject's aggregated online identity. It aims to show the observer how the Internet sees them, and in this process, show the promises and pitfalls of machines assessments of social identity.

Upon entering a name, Personas scours the web for information and attempts to characterize the person - to fit them to a predetermined set of categories that an algorithmic process created from a massive corpus of data. The computational process is visualized with each stage of the analysis, finally resulting in the presentation of a seemingly authoritative personal profile. It helped bring the concept of data portraiture to the masses with over two million uses, while validating a graphic approach towards characterization and highlighting its potential for observers to read too much into it.

In a world where fortunes are sought through data-mining vast information repositories, the computer is our indispensable but far from infallible assistant. Personas demonstrates the computer's uncanny insights and its inadvertent errors, such as the mischaracterizations caused by the inability to separate data from multiple owners of the same name. It is meant for the observer to reflect on our current and future world, where digital histories are as important if not more important than oral histories, and computational methods of condensing our digital traces are opaque and socially ignorant.

4.4 Experiment #4: Defuse

Not everyone in online public settings seeks rich discussion. Often, casual users seek to quickly scan through a forest of interaction to get a quick understanding of what the contributors are saying; a difference that should be reflected in the design. But how can this be solved as the Internet increasingly inches towards Borges' Library of Babel¹⁰ (1941)? The high connectivity of the web affords an ever-increasing number of points of view, "true" facts, "false" facts, and tangential commentary and reactions for any given situation. Current media do not sufficiently allow easy comprehension of such a large amount of data, providing little context or summary. Reverse chronological and its approximations seem to be the staple of presentation. It is difficult enough to quickly ascertain the breadth of viewpoints that exist and their thought process or validity, let alone any community-centric information about the posters and what viewpoints are typical for them. Hiltz and Turoff (1978) long ago foresaw the desire and positive possibilities that could arise from shared online dialogue across heterogeneous audiences. Now that this dream has become a reality in 2011, current asynchronous forums and commenting systems resort to paging long linear lists. With the increasing volume of opinions and article sharing, current interfaces are

¹⁰ The mythical Library of Babel contains not only every book that exists, but every book that could exist.

must rethink the list. *Defuse* attempts to do so, focusing on improving our ability to understand participants and the crowd. It uses various statistical and natural language processing methods to summarize subjects by their commenting history, and aggregates it even further for each article. Observers are given an interactive portrait of the crowd by the found demographics, that facilitates faceted drilling down into the raw comments. A data portrait of each author accompanies their messages. It demonstrates that machine learning of users' digital footprints can facilitate the social and sociological navigation of crowds. These prototypes satisfy goals and curiosities that are political and demographic in nature.

PROBLEM

It is a common goal for many to use the power of the globally connected Internet to break down traditional barriers erected between social groups in real-life in order to enable the better passage of ideas and viewpoints across society. With the recent blood soaked rise of the Arab Spring heavily fueled by social media, the awesome power of online public discourse is truly present. While longstanding dictatorships are being questioned and revolted against across the middle east, less volatile transformations and transition points are occurring in China, the US, Greece, Portugal, Spain, Brazil, India and more. State-censorship aside¹¹, if the world is going to better engage in democracy then the tools must be designed to support mass scale from the start. Defuse is one attempt at enhancing the ability to see a demographics of a crowd while simultaneously providing better cues into each participant.

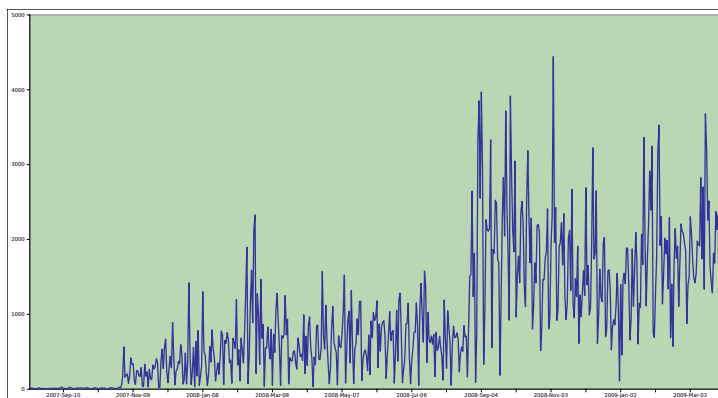


Figure 4.20. Volume of comments per day on the NYTimes.com website, from September 2007 - March 2009.

Defuse was conceived in the wake of the US presidential elections of 2008, where the average number of user comments in the NYTimes nearly doubled as the race heated up, as shown in Figure 4.20. It has since sustained the dramatic increase, indicating that online participation might be continuing to grow.

According to the Pew Internet & American Life Project (Smith et al., 2009), 15% of online citizens have

posted comments online of political nature. This should increase as the digital natives age, given 18-29 year olds participate the most. Given the massive participation in comments, it is surprising

¹¹ Many countries have attempted to prevent social networking influence in revolutions by limited access using commonly bypassed country-wide Firewalls. In one extreme, Egypt temporarily shutdown all Internet routes into the country (Williams, 2011).

to see so little change in presentation over the past decade. Even recent comment platform providers like Disqus, Intense Debate, and Echo focus their innovation on the “real-time” nature to their products rather than rethinking what participation could be, or how it might be better represented. Given popular articles can have thousands of comments, perhaps its time to rethink the experience.

Defuse is such an attempt to push in this direction. Specifically, Defuse focuses on enriching existing discussion on NYTimes.com by using the history of a user to understand themselves and the crowd around them. NYTimes.com is one of the most trafficked sites on the web, ranking #15 in the US in unique visitors (comScore, 2011). As such, its comments sections are very lively, often with thousands of comments on important articles.

Issues with the current NYTimes.com implementation

Like most comments implementations NYTimes.com presents comments in a long, linear presentation with limited filters. There are at least five problems in altering the current design: 1) there are too many comments to read, 2) individuals lack weight and perceptibility beyond their current comment, 3) the filters provided do not answer many questions, 4) no representation of the crowd and if they represent the normal NYTimes.com user, and 5) recommendations counts are not proportionally visually distinct. This section outlines these problems in more detail.

In Figure 4.21, we see the article from 2008 when Sarah Palin was introduced to the US via the Republican National Convention. In reaction to article, commenters left 2488 messages spanning 98 pages worth of text. The default sorting method is reverse chronologically, as is the standard paradigm. This format is useful for those who wish either read a few comments, or dive deeper by continuing to read more. It is a pretty good compromise between design simplicity as a 1:1 representation and yielding a representative sample. Because the time a non-threaded comment was written is mostly arbitrary with only a time zone bias, those comments read at the top will quickly approximate the actual histogram of content. Statistically speaking, for an article with 2,500 comments, a user would need to read 24 comments to have a 20% margin of error and a 95% confidence level¹². If we wanted a 5% margin of error in our perception of the comments’ gist, we would have to read 333 comments. Assuming the average interested user reads 5 comments out of the 2500, they will have a 44% margin of error. The very lazy with 2 comments

¹² The 20% error rate is an obtuse concept given it applies to the semantic abstraction performed within a human mind. It is better to think that with 20% error the user will get a decent understanding of what the comments say. These numbers come from using Cochran’s (1963) method shown below to determining sampling size given a fixed population. Assuming we would not know the variability ahead of time, we used 0.5 for both p and thus consequently q in calculating our needed comment sample size.

$$n_0 = \frac{Z^2 \cdot p \cdot q}{e^2}, \quad n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

read will suffer a 70% margin of error. To feel that one has a statistically good sense of how the community is reacting, one would need to read quite a lot of comments. It would be helpful for have a summary in some format to alleviate the burden.

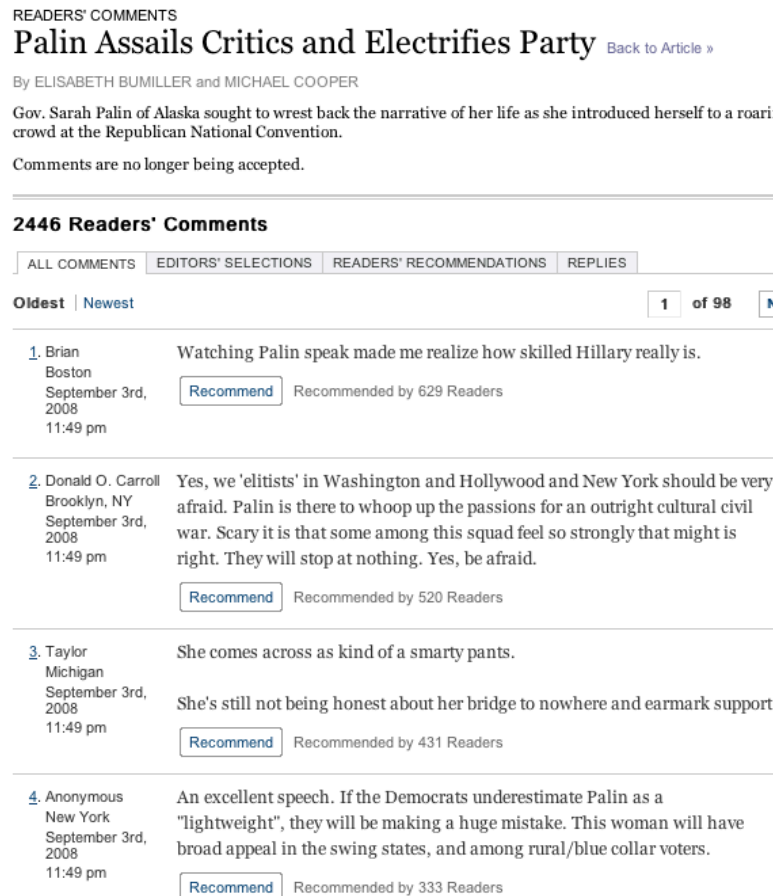


Figure 4.21. Comments from the September 3rd, 2008 NYTimes.com article regarding Sarah Palin's first speech as a White House contender.

Without an understanding of the users, we also cannot ask the interface to sort the comments in more interesting ways. Currently we can look at the most recommended articles or the editor's selections (a great method that cannot be easily replicated or automated). It is easy to get stuck on structural-level features such as number of recommendations or time because they are part of the existing model. Yet as humans we are more likely to be curious about the crowd in prototypes. What do Republicans from Iowa say? What do Red Sox fans say? What do the domain experts say?

It is difficult to currently assess the demographics of the crowd without reading many, many comments. On a site as dynamic as the New York Times, commenters could range from elected Republican officials to pregnant high schoolers in Detroit. There is no indication whether the crowds are similar across articles, or whether a single article has an unusual turnout of the

In this design, users are represented by their comment, a name, and possibly a location. Much more is possible. For example, Brian from Boston might have previously left 10,000 comments. Those messages contain a lot of information about Brian such as his opinions and concerns, sociolinguistic and psychological cues, and his ideological persuasions. Yet none of them affect his presentation in each instance today. If we could query that history to know how to judge a quirky or standout comment. We could determine if a current message is sarcastic, or if Brian is a domain expert. We might also see if Brian often is recommended within a given subject, or even more generally.

community. Users are required to read many comments over time to develop an mental image of the larger community.

Finally, the main significant quality signal -- number of recommendations -- is visually buried and not proportionally distinct. As it is so easy to recommend a comment on the New York Times, comments can receive thousands of recommendations within hours. Displayed at the bottom of the comment next to the recommend button, they are a fraction of the total comment and require the observer to read their value. As such, when simultaneously scrolling and reading, it is difficult to integrate the number of recommendations into the gestalt for upcoming comments. They are only trackable by focusing on their column, which interferes with skimming.

INSPIRATION

There have not been many examples of forums that prioritize visualizations of the crowd or its members. One example, Conversation Maps (Sack, 2001), assumes deeper threads than appear in New York Times comments and fails to include user history. Its visualizations are more exploration-driven for expert users than a queryable substrate for more likely questions. A more legible interface, Anthropomorphs, presents the crowd as a gestalt of individuals (Perry, 2004). They are a literal translation of the idea of a data body: anthropomorphic figures that depict sentiment through simple expressions, and depict their structural history on their chest (see Figure 4.22). However, in using literal representations of humans, we are more constrained in what the impressions are possible while keeping our augmentation legible. We easily misread individuals (if not always) in this form given the lack of nuance in contrast to high expectations when using the visual language of human bodies (Donath, 2001).

An interesting example of summarizing comments is Twahpic (Ramage, Dumais & Liebling, 2010), which uses topic modeling to view a given user's tweets over time in terms of *substance*, *social*, *status*, and *style*. (see Figure 4.23). To the left of a users tweets is a summarization of the tweet across these four meta-topics. To the right appears the most frequent topics across the categories, using a labeled word cloud to depict each sub-topic. The generated portrait is a bit difficult to interpret given so many of the topics are full of interstitial words: those that link semantic concepts but lack meaningful signal on their own. For example, in Figure 4.23 topics *When I*, *Positive*, and *Travel woes* are full of interstitial words like "by" and "their." Given so many patterns in language occur, we must be careful to curate either the raw representation or how we abstract from the automated clusters.

To ground curation for Defuse, we pull inspiration from sociology, in particular social scientists Bourdieu, Goffman, Simmel, and Lamonte. As discussed in Chapter 2, Simmel (1910) recognizes that we put people into human types in order to understand how to interact or interpret them. Bourdieu suggests that *habitus*, or "*the durably installed generative principle of regulated*

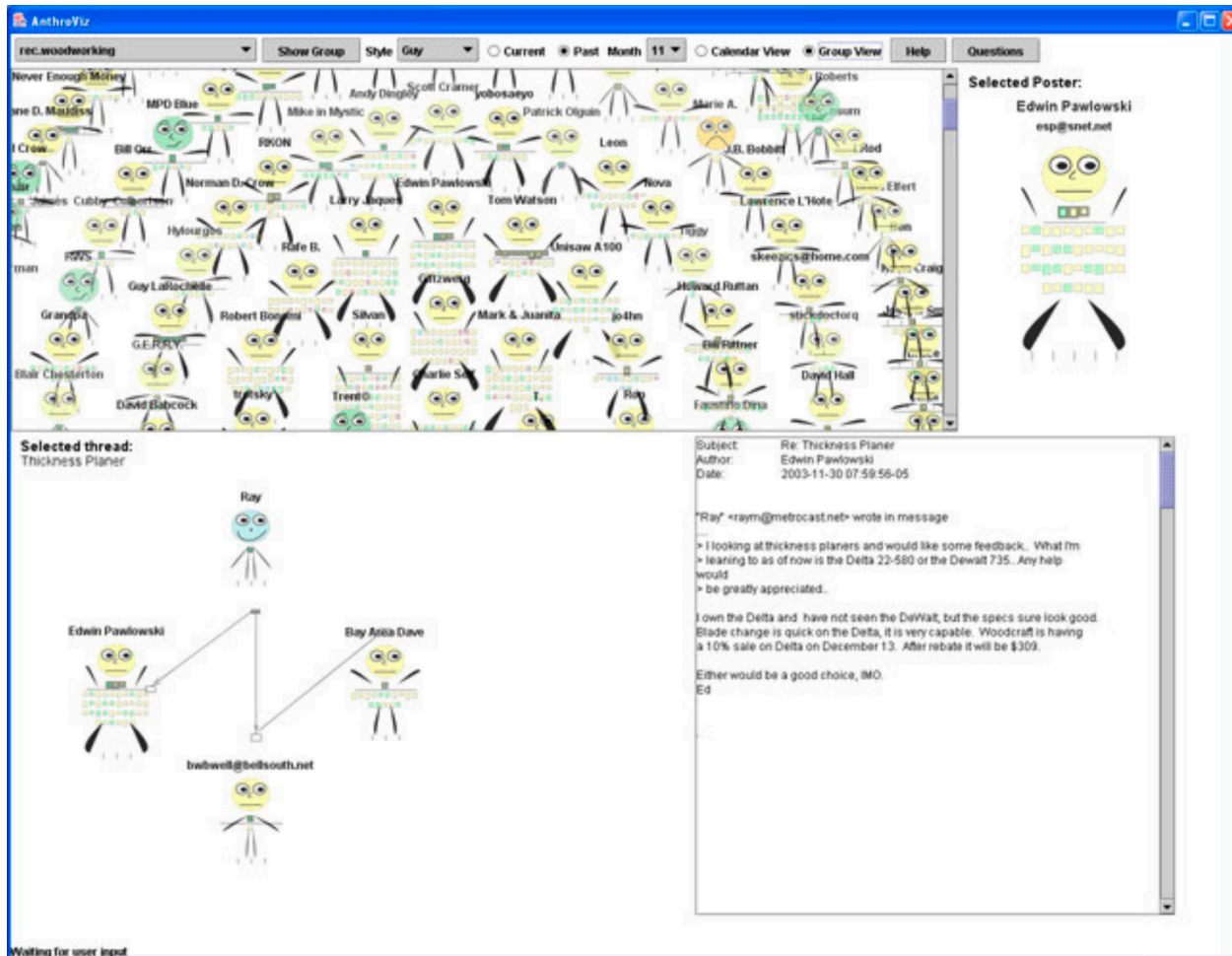


Figure 4.22. Anthropomorphs are visualizations of users that overload the human form to characterize their messages. By visualizing all the individuals together in a given communication space, a crowd gestalt is formed.

improvisations” (Bourdieu, 1977), is how we put people into prototypes. Habitus in part comes from birth, but is built up over one’s life in part by how others structure the world. In turn, we use these same structures to determine with whom we wish to interact, using hierarchies of age, wealth, power and culture. While some of these elements can be difficult to ascertain in a sparse chat space, Bourdieu found that aesthetics is able to proxy for many of these attributes, presenting more opportunities to determine the social geometry in a configuration of commenters. Lamont (1992) builds on Bourdieu, adding that symbolic boundaries such as morality and religion are equally apart of habitus and thus social practice. Many “hot button” political issues like gay marriage or medical marijuana aim directly at these common symbolic structures.

Understanding where someone falls along these axis is predictive of other relevant dimensions and thus aids in character judgement. Even though each comment may only give a slight cue into

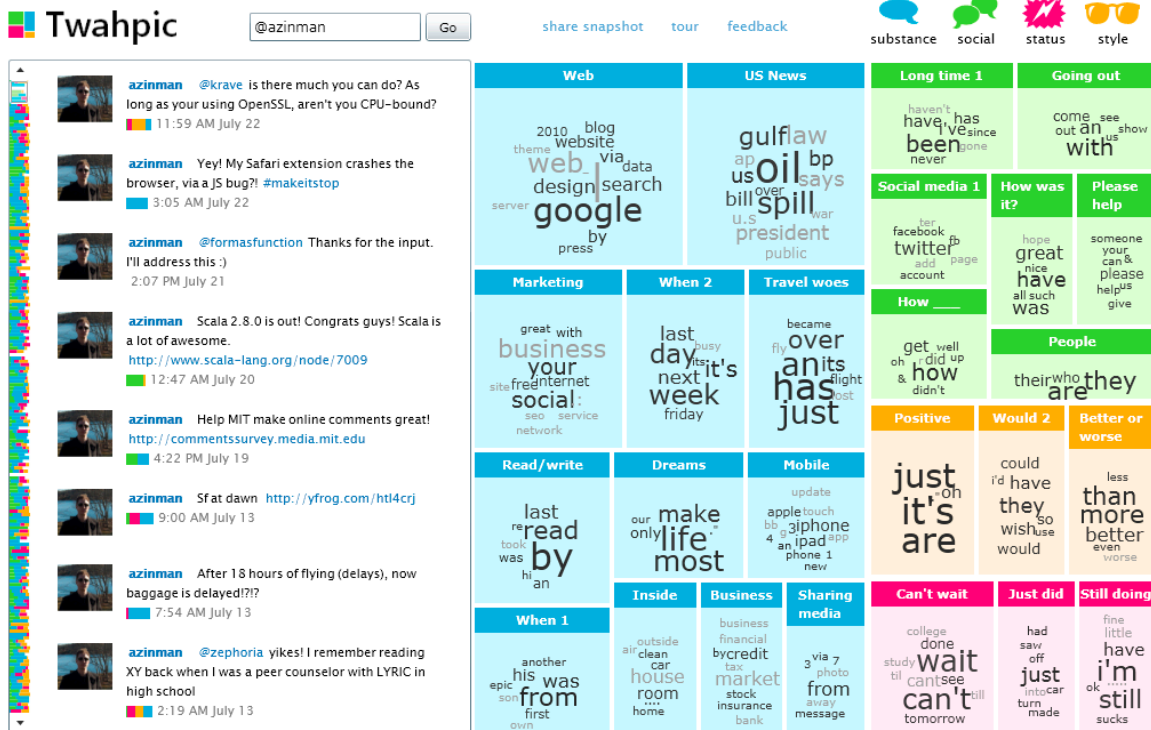


Figure 4.23. Twahpic (Ramage, Dumais & Liebling, 2010) uses semi-supervised learning to create topic models weekly for Twitter. These models are then used to create data portraits of collections of Tweets, such as the history of a given user.

the social geometry of its author; Goffman (1959) assures us if given enough recognition from a static audience, that a character or expressed identity will remain symbolically stable in its choices, thus providing a seemingly accurate view of an individual when some critical plurality is achieved. Recently this has been tested with the blog “A Gay Girl in Damascus,” where a white American male wrote from Scotland very detailed descriptions of abduction and revolution in Syria. While the fictional accounts were fraudulently portrayed and eventually reported as non-fictional, the blogger’s attempt to stay consistent in character throughout attests to the power of prototypes to be very powerful. The unique position automatically gave credibility by those who projected their habitus onto the situation, craving the anti-authoritarian narrative of an oppressed gay female American-Syrian Muslim.

SURVEY

To understand the commenting habits and desires of more digitally sophisticated users, a targeted study was performed online. It asked them questions from the perspectives of comment readers, writers, and site owners. The total population of 50 people was biased by the author, who mainly invited MIT students and those tied to the technology world. Approximately one-third of the population specialized in fields outside of computer science. They were 63% male,

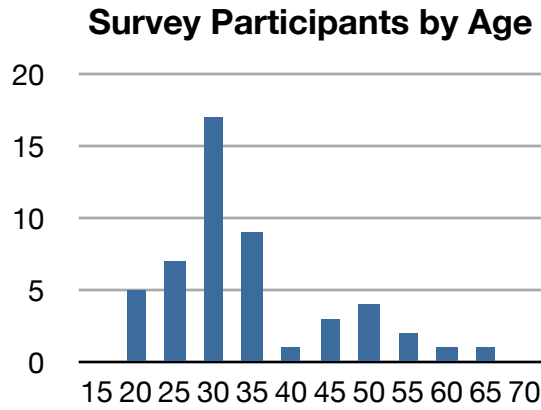


Figure 4.24. A histogram of the 50 survey participants by age.

36% female, and mostly were mostly young as shown in Figure 4.24.

For those who read or wrote comments and product reviews, they were asked a wide variety of questions about general and site-specific usage patterns, organized by the following types of sites: aggregators (e.g. Reddit), major blogs (e.g. BoingBoing, Politico), other smaller blogs, media-oriented sites (e.g. YouTube, SoundCloud), news sites (e.g. New York Times), product reviews (e.g. Amazon), social media (e.g. Facebook), or other types of sites. If they owned their own site, they were asked about user participation and their goals. The raw results to this

survey can be found in Appendix A. Here we present the main takeaways:

Users would like some method to visualize and organize the comments as a whole. They would prefer that over more manual techniques like tagging of identities to improve comments. As such, 56% would like comments to be grouped by how humans think. Most would like crowds to achieve a consensus through the medium, but most do not believe that is possible. Respondents mostly comment when they have something unique to say (19%), determined by often reading at least half the comments (33%). When they comment, 73% spent more than 3 minutes crafting their comments, suggesting that they are not as low-cost for this crowd as one might often think of Internet comments. The group as a whole were heavy participants, as 63% always or often read comments, 75% spend more than three minutes reading, and a power user 25% spend 15 minutes to 1 hour reading. When asked if they participate because of diversity, participants generally say no; political diversity had the most draw. However, when asked which types of diversity are important, socioeconomic, geographical, and educational types top the list. Respondents do care about the reputation of their account (55%), and have stopped themselves from writing a comment (65%). Some (20%) even attempt to improve their offline reputation through participation online.

DESIGN

The final version of Defuse is quite different from where it started. It evolved over four major iterations, changing in concept and approach in a way that echoes the larger trend of starting with 1:1 mappings and evolving towards abstraction. We now describe each version, and in turn, the research arc towards finding the proper semantic units to represent commenters.

Version 1

Defuse was first conceived to give person-centric views of comments as a reaction against the standard reverse chronological paradigm. The notion of a person-centric view remained in some capacity through all iterations. A person-centric view can still look at an article's comments as a collection as done on the NYTimes.com, but like Anthropomorphs we focused on making the comment a second step to beyond choosing the person with the most desirable representation. Unlike Anthropomorphs' usage of the human body, Defuse attempted to employ a minimalist design language on the representation on the heels of Personas' legibility.

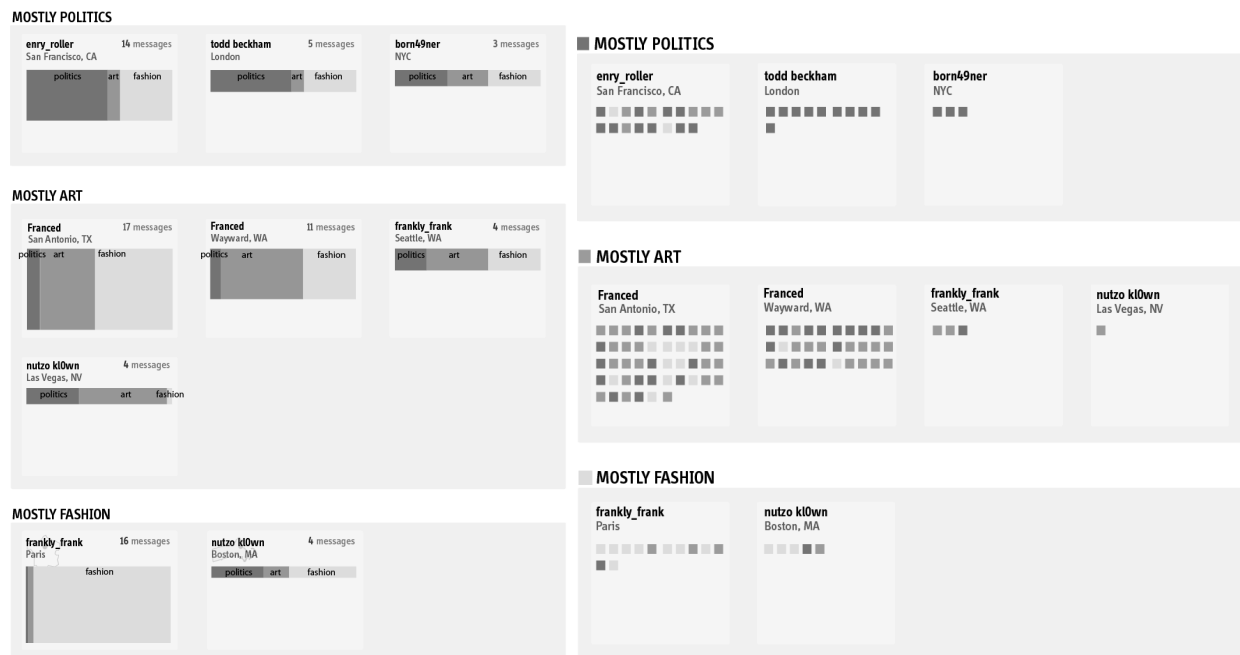


Figure 4.25. Early sketches of Defuse. The sketching process started by visualizing the structural habits of users. In both versions, users are clustered by the newspaper section they frequent the most. On the left, the first concept examined representations of posters in topic and total volume using an extension Personas' visual language. On the right, the representation breaks from the continuous bars, freeing each message to be individually annotated.

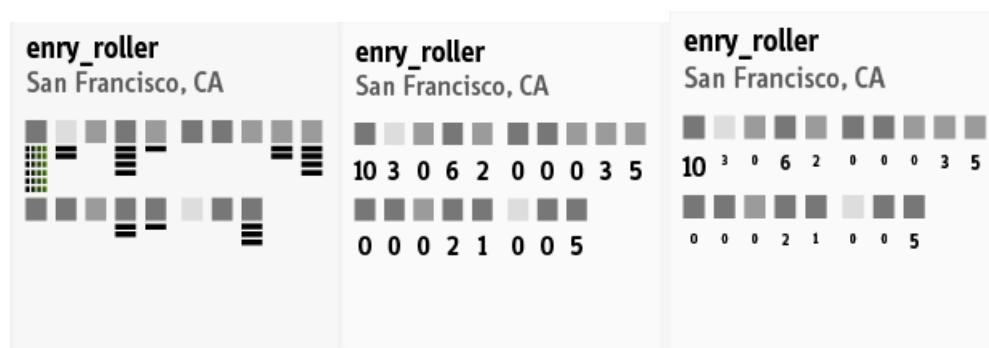


Figure 4.26. Various attempts to represent the number of recommendations a message received.

Starting with Personas as motivation, Defuse first sought to represent individual primarily by the section of the newspaper they comment on the most (representing prototype and habitus), and then through their posting history at a structural level. As seen in Figure 4.25, we progress from a flawed Personas' restyling towards a more literal representation where each message has its own square in something we refer to as an *author view*. This was done given the perception of volume is non-linear (Mackinlay, 1986), and did not offer the ability to further distinguish between the messages. On the right, we see a cleaner representation where each message stands on its own. It is easy to estimate overall volume, as well as to judge the section heterogeneity. Figure 4.26 depicts sketches that annotate messages in proportion to the number of recommendations received. On the left, visual tick marks provide a better gestalt than requiring the user to read individual numbers.

Figure 4.28 shows a screenshot from the working first version using real data. After picking a given article, we can scan for individuals we might find interesting, hovering over them to reveal their entire past history or clicking to reveal their comments (see Figure 4.29). Comments have stars that scale to the number of recommendations to provide a consistent graphic violator who's value can be easily estimated roughly by size. Comments are revealed in a focus+context, motivated by Google Reader. The sizes dynamically scale from small towards large for the

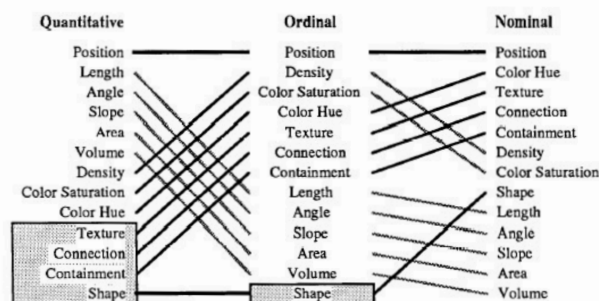
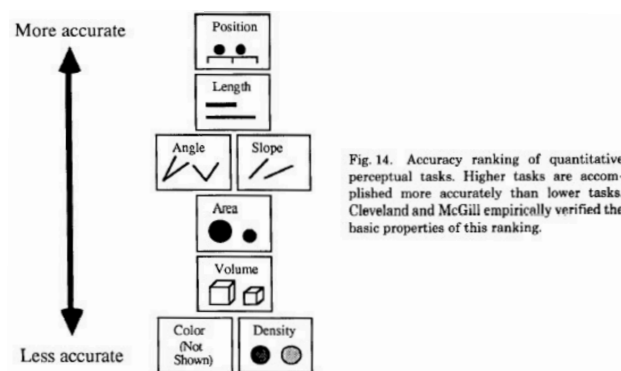


Fig. 15. Ranking of perceptual tasks. The tasks shown in the gray boxes are not relevant to these types of data.

Figure 4.27. Mackinlay's (1986) ranking of visual attributes to aid in perceptual tasks.

“current” comment in the middle. By focusing on one comment at a time, we can time their visit length so as to gain the implicit quality signal ($f(\text{time}) / g(\text{comment length})$).

In its simplistic 1:1 mappings, this version remains very objective in its representation. The structural qualities partially answer several useful questions: 1) which users are recommended in general, or within a given section, 2) what are the crowd demographics (using section clustering to prototype persona), and 3) who the main contributors are, and which comments are the author’s first, 4) how diverse are an author’s posts (determined by section). Yet there are many useful questions it cannot answer. What do Republicans from Iowa think? What about Red Sox fans? What is the general mood in the crowd? We need to start exposing and characterizing the raw text to answer these questions and more.

Figure 4.30 shows the sketches for the intended follow up to *Version 1* that never were built. The comment-centric view was conceptually scrapped in favor of retaining a person-centric perspective to be consistent with the thesis. However, the use of sentiment analysis, topic modeling, Meme-Tracker (Leskovec, Backstrom & Kleinberg, 2009), and geography were all employed or attempted in later versions.

The visualization was also adopted to use product reviews on BestBuy.com, but the sparsity of the reviews made it difficult to extract meaning. This version and all versions of Defuse were built with the assumption that a critical density of information will eventually be achieved universally, which is when tools like Defuse will significantly enhance engagement online.

EDUCATION

JANUARY

113

COMMENTS BY FRANCD

◀ Back to everyone's comments

MOUSE

click a box

use the scroll bar

KEYBOARD

j goes down

k goes up

Franced

San Antonio, TX

JANUARY



FEBRUARY



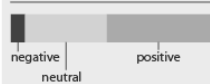
TOPICS

[?]



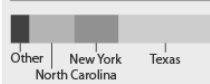
SENTIMENT ANALYSIS

[?]



GEOGRAPHICAL PREDICTION

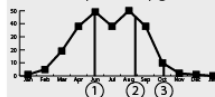
[?]



MEMETRACKER

[?]

PHRASE: "lipstick on a pig"



- ① ..they'd slap lipstick on a pig around..
- ② dont put lipstick on a pig and call it..
- ③ ..its just lipstick on a pig as far as im..

January 3, 2009 12:58pm

+5 recommendations

We need to negotiate a strategic agreement with Iran, the HRCOCS and Israel's are doing the same thing they did with Iraq, waiting until we're in a bad spot then attacking the US or its neighbors. Israel should not use force to solve problems as they are too big to be solved by force. If they do, we should be in a bad spot then.

January 1, 2009 3:02pm

+5 recommendations

I think this is a fine thing, but why aren't we involving the kids? The kids should be at least assisting the cooks. Chopping, cutting, clearing the food. Cleaning the kitchen, the Cafeteria. Learn by doing, and then doing it over again.

January 2, 2009 6:02pm

I think this is a fine thing, but why aren't we involving the kids? The kids should be at least assisting the cooks. Chopping, cutting, clearing the food. Cleaning the kitchen, the Cafeteria. Learn by doing, and then doing it over again.

January 3, 2009 12:01am

+9 recommendations

1) Something strange that the US News Media is misrepresenting is that Iran has NOT "buried the new facility inside of a mountain". If you look at the site with Google map Terrain feature, you see that it is under a small hill about 60 meters high. That is pretty shallow protection.

2) As FAS and others have pointed out, if you want to protect an illegal bombmaking facility, you need to tunnel under a granite mountain at the base -- so that you have 1000 feet of granite overhead to protect you from a nuclear strike. There are plenty of places in Iran that would provide far better protection if it was going to pursue an illegal program.

3) I agree with those above who say that if the US government is serious about nonproliferation, it needs to rein in the Israelis. Bibi is holding a knife to the throat of every Middle Eastern country. Of course, if we had Iranian billionaires in this country who purchased dual citizenship like Haim Saban and dumped \$15 Million into the Democratic National Committee's coffers, there probably would not be any concerns expressed about Iran's intentions. Maybe Iran should have its expatriates here form an American Iranian Public Affairs Committee (AIPAC).

reply share bookmark

disapprove | recommend

January 5, 2009 12:58pm

I think this is a fine thing, but why aren't we involving the kids? The kids should be at least assisting the cooks. Chopping, cutting, clearing the food. Cleaning the kitchen, the Cafeteria. Learn by doing, and then doing it over again.

January 18, 2009 2:38pm

Thanks for a good retrospective and objective article. Please remember long term history as well. The Lone Danger (Lone Wolf) isolated emotion in Iraq with his way ignorant - Acts of evil speech which degraded the Iranian public's intelligence to the point where they elected a Hardliner. Now we are suffering the consequences. It's because he made a power after a presidential election. Thank you for being a beacon of reason.

We're sick and tired of this crap. We're mongering world. We're corrupt. Please declare PEACE!
Rethink everything.

January 20, 2009 1:58pm

+5 recommendations

I think this is a fine thing, but why aren't we involving the kids? The kids should be at least assisting the cooks. Chopping, cutting, clearing the food. Cleaning the kitchen, the Cafeteria. Learn by doing, and then doing it over again.

(a)

ARTICLE COMMENTS

◀ Back to article viewer

FILTERS

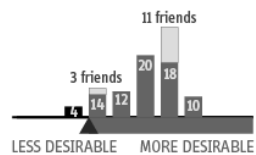
SPECIAL STATUS

FRIENDS

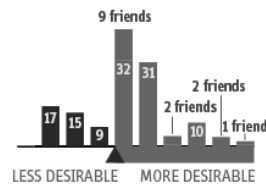
IGNORED

FOLLOWED

REPUTATION [expand]



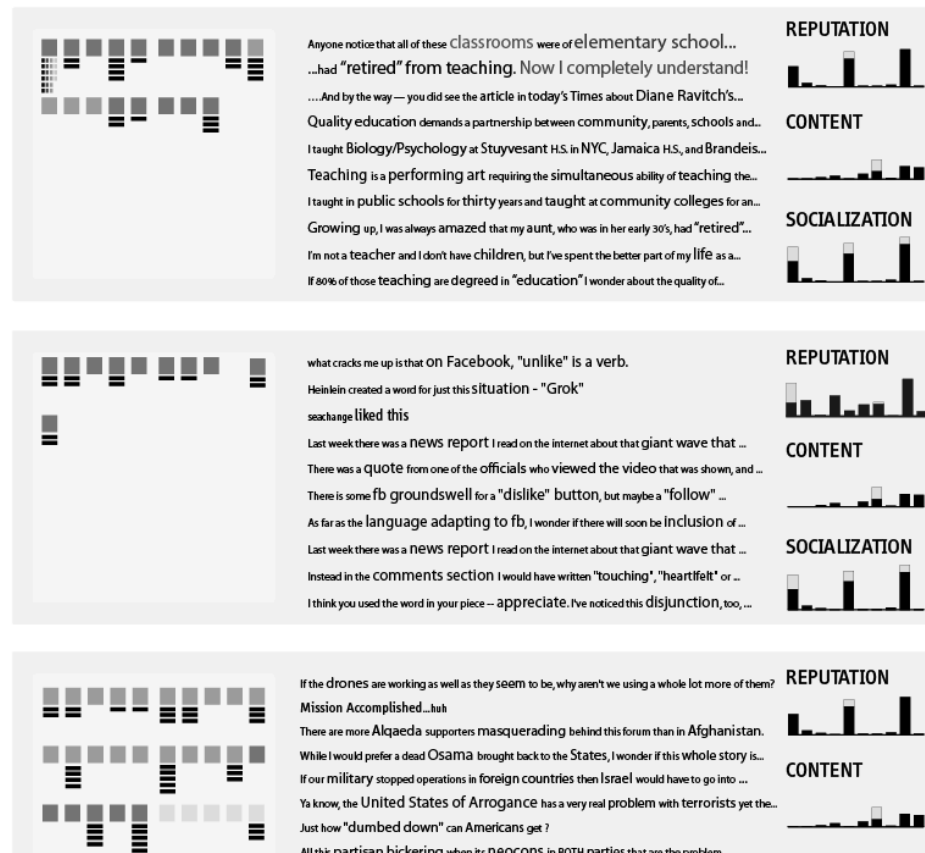
CONTENT [expand]



SOCIALIZATION [expand]



COMMENT CLUSTERS



(b)

Figure 4.30. Sketches for the next iteration which were scrapped. They (a) expanded the use of machine learning to summarize subjects, as well as (b) creating a comment-centric interface that emphasized faceted filtering across various signals.

Version 2

The next iteration first started as trying to port *Version 1* into HTML5 from Flash, it quickly started to diverge. *Version 2* started to focus on augmenting and analyzing the comments to summarize the crowd. *Version 1* relied on the user to mentally form a gestalt, where as *Version 2* turned towards statistics to create the data portrait.

Figure 4.31 shows screenshots of *Version 2*. On the left top, we can see an area dedicated to summarizing the commenters, split from the article's comments and author views. In a Personastyle we see a histogram in which newspaper sections the crowd posts. Below that lies a histogram of the crowd's word characteristics across all comments using vocabulary analysis, which looks at the ratio between words commonly known by third grade (also known as the General Service List (West, 1953)), words more likely to be found in an academic paper (Coxhead, 2000), words used

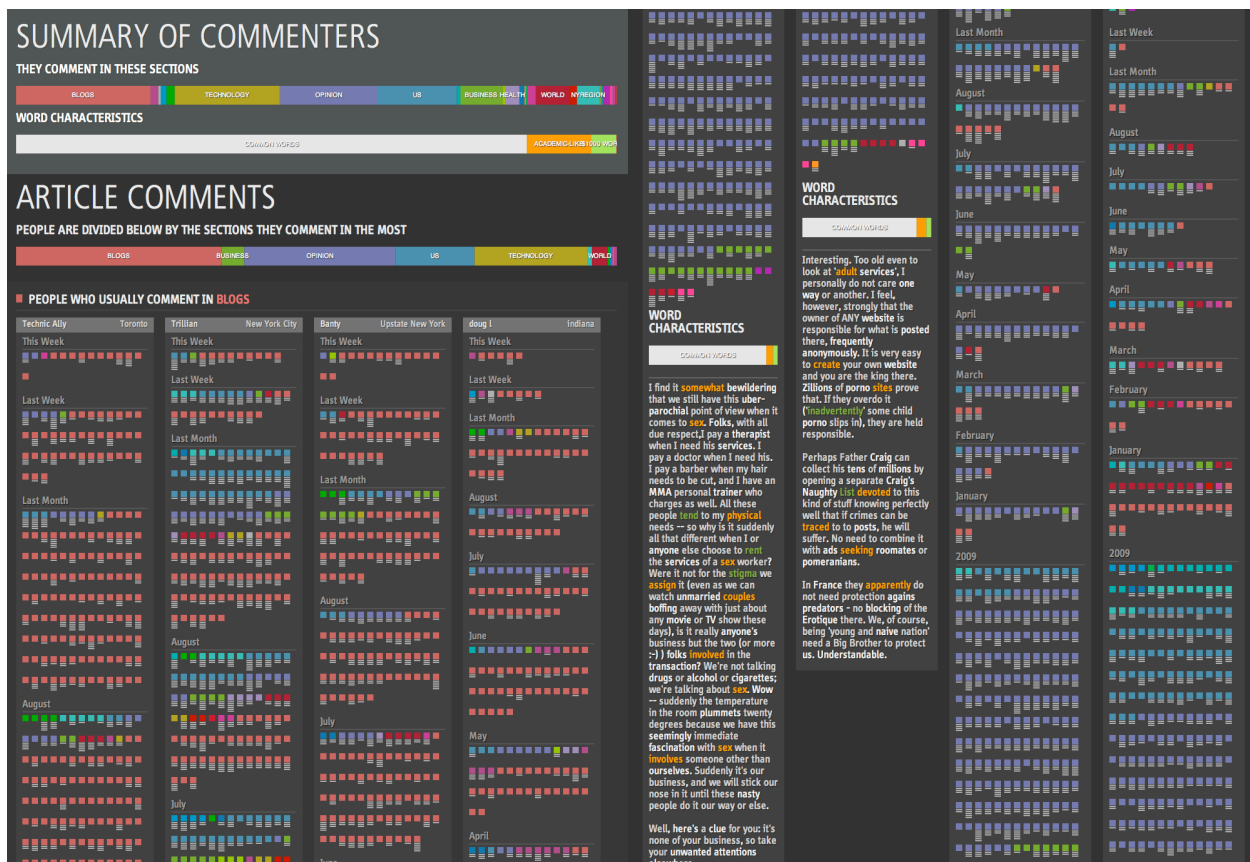


Figure 4.31. Screenshots of *Version 2*, where the picture on the right is a vertical continuation of the left. We start with a summary of the commenters by newspaper section and word characteristics (classifying vocabulary used). Below we have expanded author views similar to *Version 1*, combined with the comment itself colored by its word characteristics.

on past GRE vocabulary exams, and profanity compiled from various Internet-accessible NLP lists. These choices were not meant to be exhaustive but outline the possibility.

The second half contains the article's comments preceded by the author view of the poster. The author view is similar to *Version 1*, however it integrates the comments directly into the visualization to avoid further clicking. This was done to simply the browsing experience. The comments were further annotated using color according to their vocabulary category membership.

Summarizing crowds by newspaper section was successful: articles each had wildly divergent signatures, principally determined by the article's section of the newspaper. However, the word characteristics proved to be less useful in informal discussions with fellow researchers and lab sponsors. While the lists themselves could have had more meaning, it was not clear what purpose it could serve, and if the current representation was the best to match the analysis method.

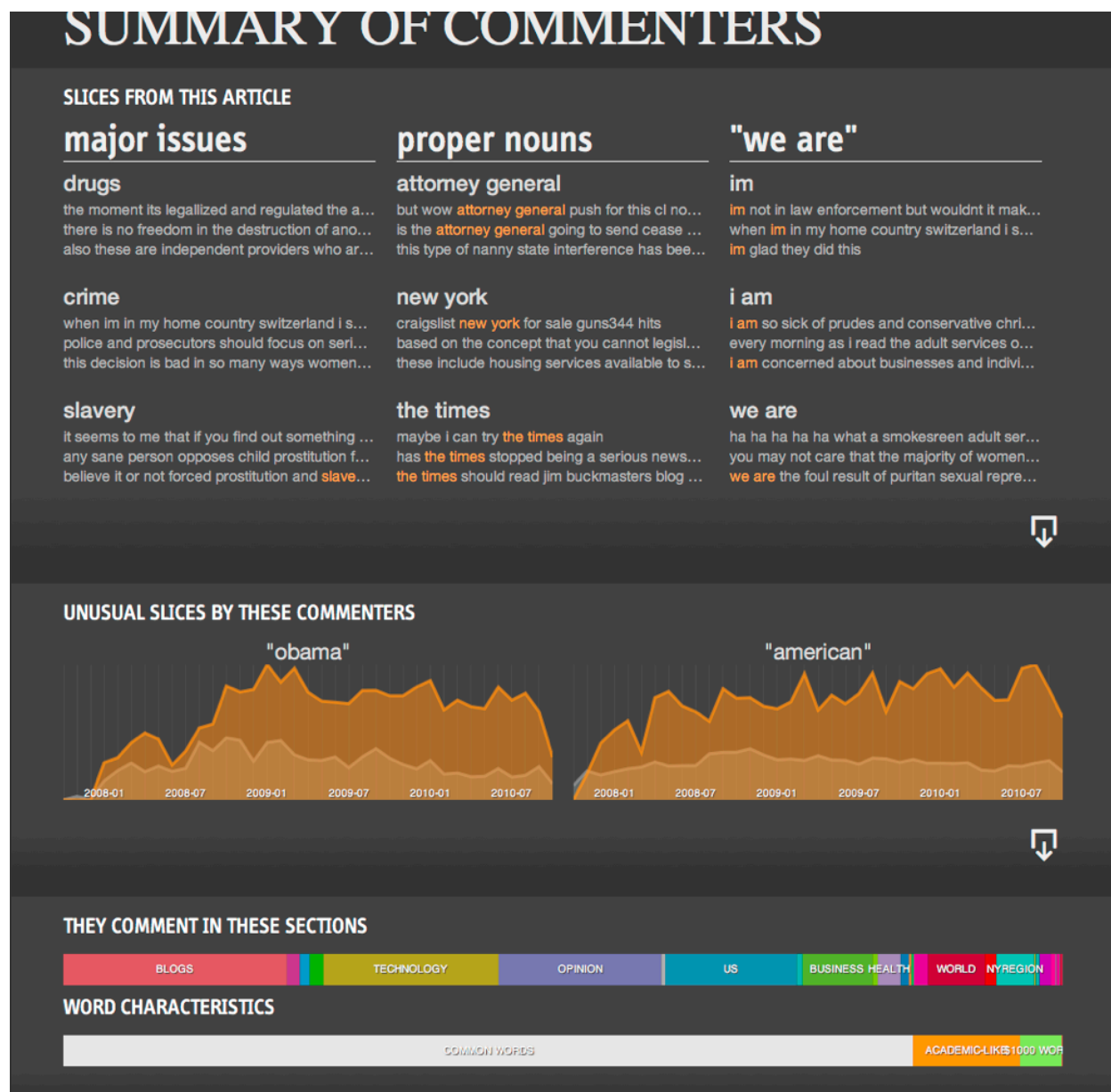


Figure 4.32. *Version 3* tries to dig deeper in the data to improve the summarization at the top.

Version 3

Seeking richer summarization into the semantic behaviors of the crowd, *Version 3* sought to further push the natural language-based character surfacing while continuing to stay objective to the data. It expanded upon the design of *Version 2*, but focused on the summarization. To find interesting aspects about people, deeper vocabulary sets were created: cultural-figures and references using Freebase (after much data cleanup), Fortune 500 companies, “*we are*” versus “*you are*” Personas-esque statements, divisive political issues (e.g. gay marriage, abortion, drugs), locations, memes from Meme-Tracker’s dataset (Leskovec, Backstrom & Kleinberg, 2009), and

political party and ideologies. Each comment in the corpus was tallied against 45,000 hierarchically-organized keywords and phrases in total, or *slices*.

The metaphor for *Version 3* was that of a party. When one enters a party full of strangers, one looks around the room trying to see what the crowd has in common, as well as how the crowd differs against one's background models of society. Figure 4.32 shows these party strategies employed at the top, where snippets from the article's comments are aggregated and ordered by the most popular slices. Below these lies a time-based visualization which attempts to show how this crowd differs from the background model. Each slice has its own associated time-based histogram of usage per person, per article, and for the background model, the entire NYTimes.com. The visualized slices are chosen based upon which slices differ significantly from the NYTimes.com on average, whether it is a significant increase or decrease in usage.

The figure shows comments from the September 4, 2010 article entitled "Craigslist Blocks Access to 'Adult Services' Pages." In Figure 4.33, which shows an expanded view of the slices at the top, we can see some slices that make sense in the context, such as *drugs* or *crime*. Other slices, such as *i am* or *we are*, are jarring in that they are cognitively disparate with the article itself and the other slices. It is also difficult to parse the text, as it is not clear why a snippet is there and what the slice means, where the slices come from, and the 1:1 mapping leaves a high cognitive overhead to find common value in the selection. The time-based histograms, on the other hand, are less words to process, but suffer their own unique problems. The sparsity and irregularity of participation means that normalization math used cannot ever be correct, simply because we cannot visualize users who have participated at different times and quantities in the same averaging figure as we attempted.

Version 4 (final)

The primary issue with *Version 3* was that it presumed being objective meant you could show possibly interesting data to the user, and require them do all the cognitive work of extracting meaning. The word choices cannot be interesting in a global sense, but should instead focus on solving actual user goals. While 45,000 slices will be useful for a variety of goals, knowing which the user is looking for will be impossible. Even if found, we still are requiring the user to further synthesize the data, and unless we show the raw data for everyone, we cannot expect to form an accurate estimation of the crowd. Instead, *Version 4* shifts to focusing on collapsing the data behind upon high-level semantic units that can directly answer questions the user might already have of the data, in addition to informing them the various biases that exist in the crowd.

To organize the crowd, we pull inspiration from sociology again. We choose five high level dimensions: Social (Goffman), Political (Lamont), Cultural (Bourdieu), Linguistic (Bourdieu and Hudson), and Economic (Bourdieu and Weber). They were chosen to be mostly orthogonal

dimensions, but there is some overlap. These dimensions are guides for the user to ask the following types of questions:

Social: How do the commenters interact with others? Are they good community members?

Political: What are the political viewpoints and concerns of the commenters?

Cultural: Where do they fit within society? What are their influences? What topics concern them?

Linguistic: How do they speak? What tone of voice do they use?

Economic: What is their economic background? How conspicuous is their spending? Might they be peacocking?

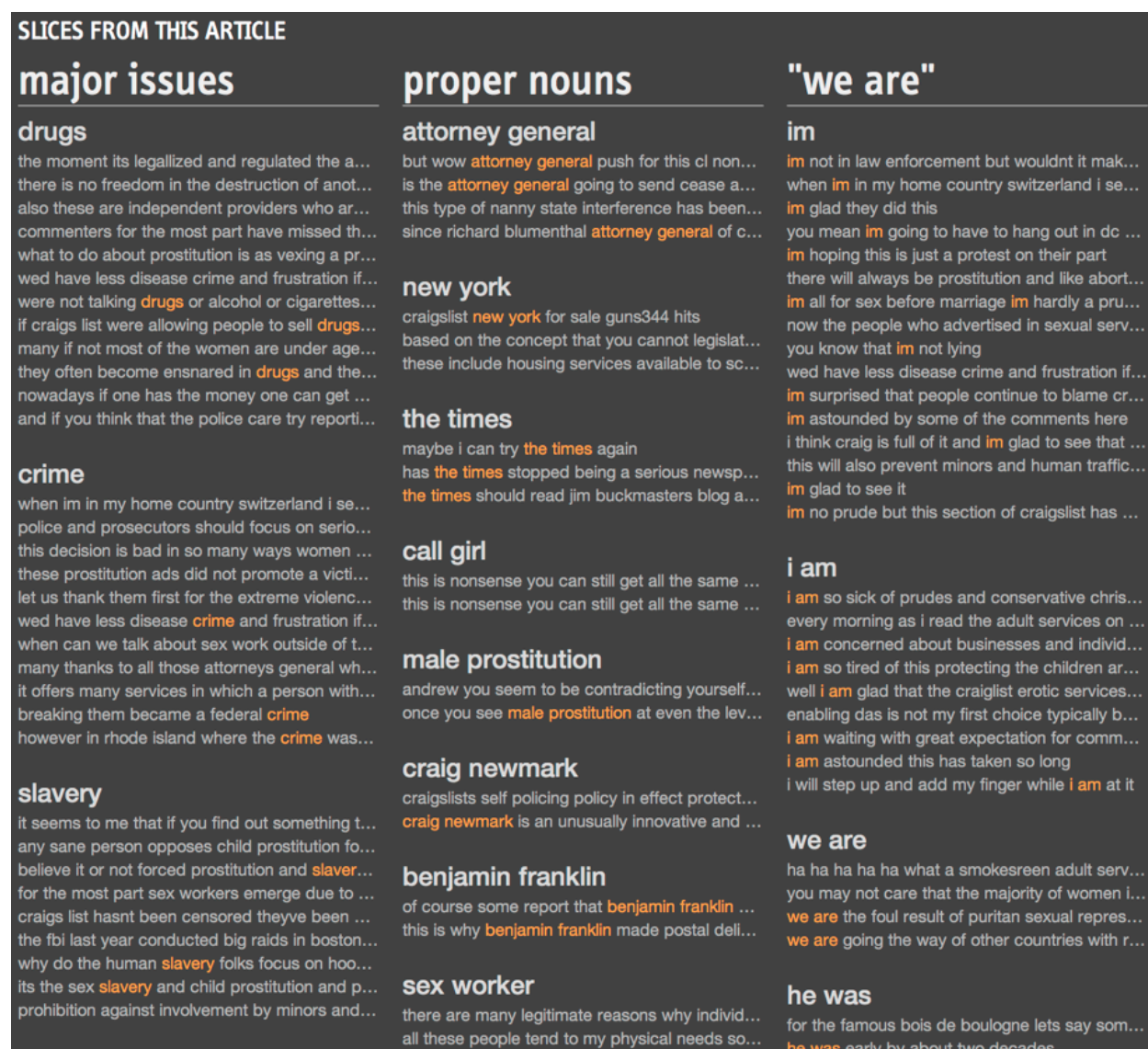


Figure 4.33. Expanded view of the slices on the September 4, 2010 article “Craigslist Blocks Access to ‘Adult Services’ Pages.” Approximately 45,000 slices were generated from a variety of data sources, and used to find deeper semantic concepts in the data.

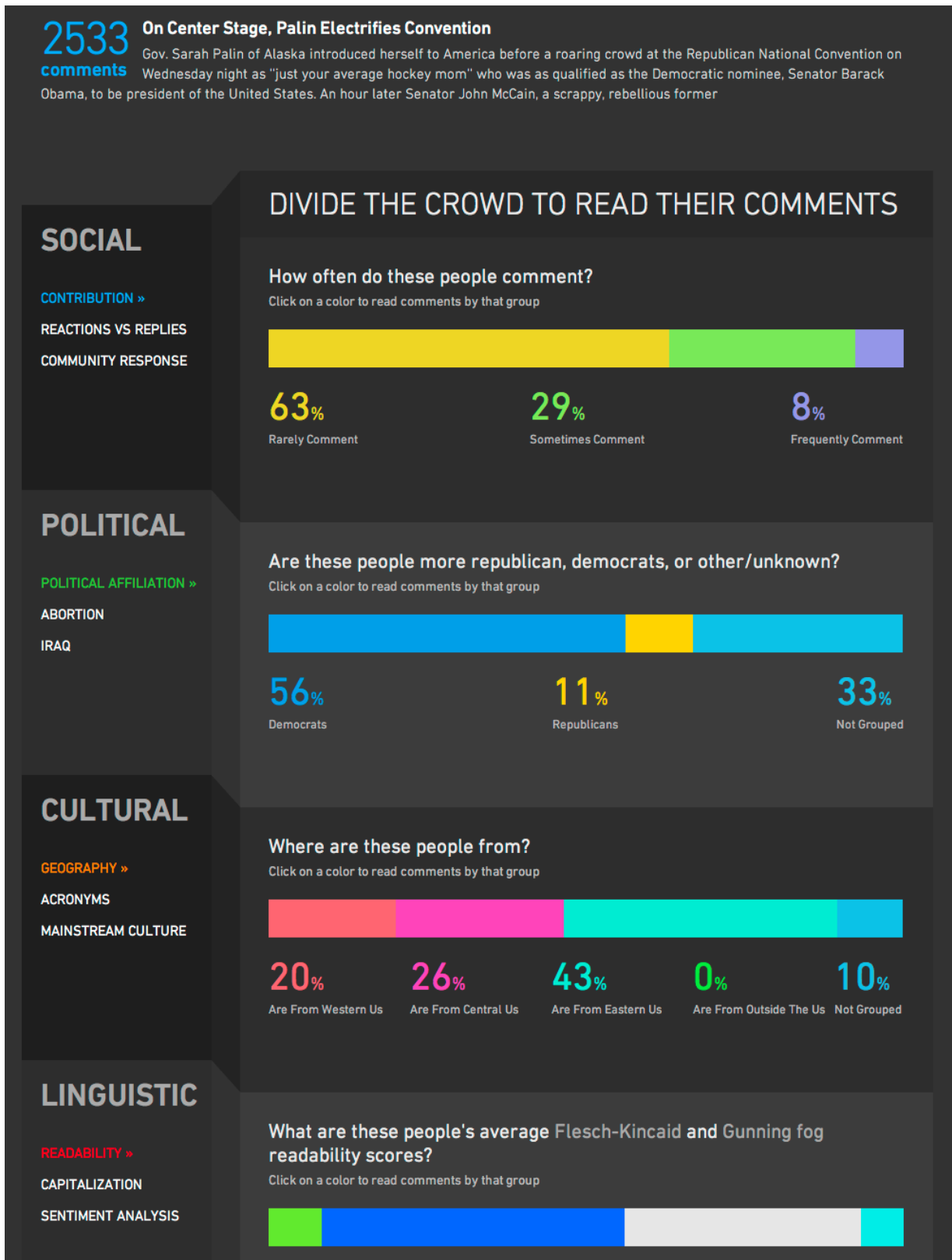


Figure 4.34. The final version of Defuse, viewing the Palin article showed previously.

They are not an idealist list that will exhaustively assist all users, but represent how society and communities already fragment themselves. We must recognize the bias as the data portraiture artist in choosing these fundamental dimensions. The Economic dimension is largely left to convey the idea, as the comment corpus contains too little information to be able to accurately guess an author's financial condition. Future work should examine linking financial sites like Mint.com with the data to be able to provide empirically-driven aggregations.

For each dimension, several filters were created that fragment the users by applying a heuristic to their past posts. For example, heuristics range from simple posting frequency statistics to classifiers that determine political affiliation. The filters were chosen as a combination between what was easily possible, what was interesting, and what best showcased the theory. Future work should include surveys to determine the ideal filters; it was felt a non-ideal proof-of-concept was suitable for research purposes to first validate the experience and interface.

Each filter breaks down the crowd into a series of expandable buckets. Clicking on any one bucket reveals the article commenters that belong to that subset. Like *Version 2* and *Version 3*, the author view visualization joins the article's comment. However, the visualization has become more compact by using opacity to represent recommendation and removing the grid spacing, aggregates by newspaper section versus time, and sorts by recommendation to provide a better gestalt. Joining the historical view is the result of the various filters as applied to the author and broken down by the five key dimensions.

IMPLEMENTATION

Defuse changed significantly in both backend and frontend over the four major iterations. In *Version 1*, it was started as an Adobe Flex/Flash project that received its data in JSON format over HTTP to a Python/CherryPy/PostgreSQL backend. All pre-processing of the data was performed in Python and written out in a de-normalized form to the database. The NYTimes.com data itself was collected using their public API, building a database of 2,237,679 comments from the years 2007 through 2010. There were many issues with speed in using native the native Flex data bindings for the visualizations, resulting in customized drawing routines that intelligently cached data to later lazily render.

In deciding to move to HTML/Javascript/CSS for *Version 2*, the slow processing speed of Python motivated a subsequent backend change to the Java Virtual Machine. Using Jetty and Jersey to implement REST calls over HTTP, the custom pre-processing and web code was built using Scala to retain Python-like programmer productivity while retaining the speed advantages of static typing on the JVM. The PostgreSQL server was retained while the rest of the front-end was reported on top of jQuery in JavaScript. All rendering was performed client-side in

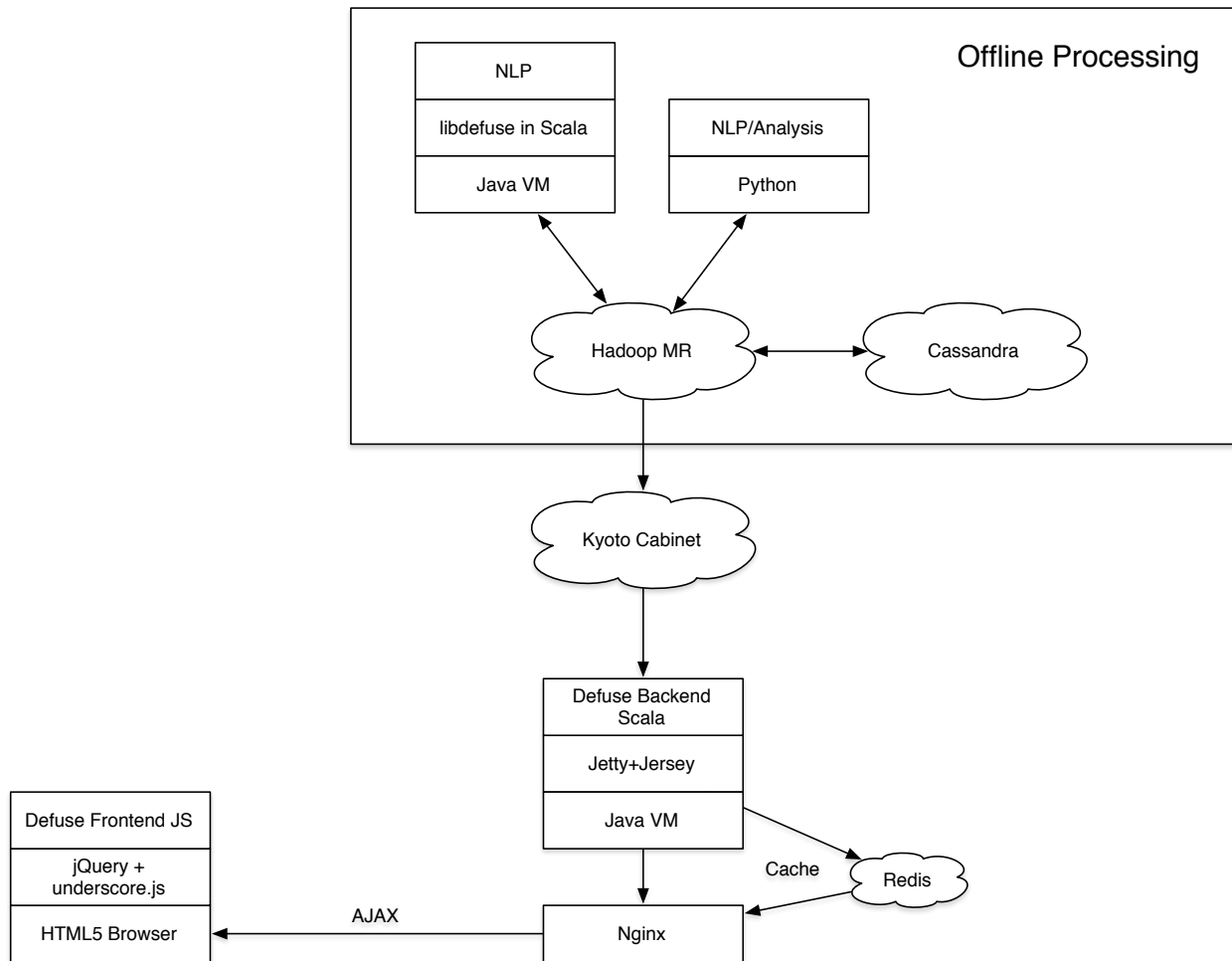


Figure 4.36. A systems description of the final version of Defuse. Web browser clients employ a RESTful connection to the backend, which serves from pre-computed caches.

Javascript in reaction to the same JSON as used in *Version 1*. The author view visualizations were implemented in pure HTML/CSS with colored DIVs using the Underscore.js templating library. Defuse continued to use the Jetty/Jersey/Scala backend and jQuery/Underscore.js frontend base throughout. The Personas-style bars were rendered using the Javascript-based visualization framework Protovis, which was later replaced by standard HTML DIVs.

Due to the significant increase in data processing and manipulation for *Version 3* and its surrounding NLP experiments, PostgreSQL was replaced by Apache Cassandra. This made it easier to dynamically add columns to comments with the results of analysis, as well as perform processing in stages using Apache Hadoop's implementation of MapReduce (Dean & Ghemawat, 2008). NLP pipelines were built that searched for and counted n -grams found in the vocabulary sets, dumping the results into author and article-specific rows in Cassandra. FreeBase's cultural list created by first filtering their public domain database dump for categories of people and locations, then against garbage and non-English entities, sorted by frequency in

the NYTimes.com corpus, and finally manually edited to clean entries. Other NLP techniques were attempted, including finding co-locations, significant phrases (n -grams), building corpus and personal topic models, within-article agglomerative and k -NN clustering using cosine distance of $tf-idf$ vectors, and using 3rd party services such as OpenCalais. These are discussed in the machine learning chapter.

While MapReduce and Cassandra worked well at first, the dynamic schemas and high-overhead costs made writing new filters difficult. *Version 4* moved away from a dependency on Cassandra, instead writing many Kyoto Cabinet key-value databases to disk in both normalized and de-normalized forms. Most of the filters built across the five dimensions were small lines in Python that ran against a master Kyoto Cabinet database. These were stored using the author as the key, composed for an article on demand through an article-author Kyoto Cabinet index. Author views are lazily loaded via AJAX depending on which authors the bucket reveals. Nginx was used to deliver static files, and Redis was placed between the JVM and Nginx to cache all AJAX compositing requests.

The source code of Defuse has been publicly accessible using Bitbucket.org throughout development, and is licensed using the GNU Affero General Public License (AGPL).

DISCUSSION

Creating Defuse was the most difficult of the demonstrated experiments. It represents two years of work in trying to understand commenting communities using machine learning and visualization. While briefly Defuse attempted to use the BestBuy.com product review dataset, sparsity issues proved challenging to work with, and as such, it kept focus on solely using the NYTimes.com commenting corpus.

While Defuse started by attempting to stay as objective and structural as possible, it eventually was found that more abstraction was needed to communicate at a semantic level that is useful to users. While the abstraction becomes more subjective and erroneous as classifiers are introduced, the trajectory is ultimately the right one because any other approach either requires too much user effort or cannot answer useful or difficult questions. Classifiers will get better, as will user expectations and influences on the types of filters and prototypes achieved. We believe the approach is a fresh take on the typical sorted lists found elsewhere: subdivide the crowd on habitus rather than ordering them by structural features.

The final iteration is not perfect, but the power of commenting history is successfully illustrated. Defuse already lets us see trends that were otherwise opaque, finding interesting ways to pull apart viewpoints or remove the noise. It is one of the few attempts to take thousands of comments and meaningfully organize them beyond what is contained in the comment alone. It

also showcases the power of a strong history to provide context and weight to comments that would otherwise stand alone. We can judge individuals by community merit, posting habits, as well as various attributes inferred from the comments themselves. In a world that is increasing digital, it gives the ability to dive deeper into a communication act -- this is important considering that history serves to provide context in a world full of astroturfing¹³ and other malicious manipulating intentions.

There are many next steps to improve Defuse. First, it might be useful to explore word-cloud visualizations for authors to summarize individuals or articles intelligently (i.e. not a straight-forward implementation, but rather expanding upon the existing Calais-provided vocabulary to give interesting and worthwhile slices). Further developing this concept, these word clouds might be overlaid on top of the author view to reveal semantic trends inside an author's comments.

A large flaw of Defuse is that it fails to deeply explain how the data was abstracted and why those dimensions were chosen. For example, how much a user talks about abortion appears regardless of an article's context, which is not always useful. It probably is not relevant to an article on the Super Bowl. Further, the data presented is somewhat difficult to interpret given the large number of absolute values shown. Pairing each visualization with one of the NYTimes.com community as a whole would help provide a normative basis for interpretation. Similarly, expanding beyond the NYTimes.com for community norms could help contextualize the larger community within the context of the rest of the web, showing for example FoxNews.com norms along side the NYTimes.com.

Currently Defuse is a stand-alone system that is not integrated into any existing commenting systems. Web browser extensions could be built to replace the existing NYTimes.com platform with Defuse. Even more exciting would be integration into the article itself; what does it look like to have commentary usefully annotated? DisputeFinder (Ennals et al., 2009) is one example of weaving outside sources to help validate or cast doubt on claims made in the article, but it utilizes more structured sources such as Wikipedia. Understanding what cross-references are meaningful to show is a difficult outstanding challenge.

Another outstanding challenge is to summarize comments directly rather than relying on demographics to filter the dataset. In its ideal form, summarization would go beyond keyword extraction to generate high-level summaries of the crowd and what main viewpoints or concerns are being expressed. As NLP and AI are a long way off from making this a reality, some hybrid human-machine interface might make this a plausible reality. Without a new interface, it is already possible to classify subjective versus objective statements and other qualities of prose and

¹³ Astroturfing is a term to describe commercial or political interests posing as grass-roots advocacy online, posting anonymously or using fake accounts to further an agenda without proper attribution to the campaign.

style, providing some basis to help sort through the comments. As every message is has varying quality and purpose, why do we present them all the same? It is likely that we could better differentiate insightful commentary from political graffiti. If automated techniques could accurately partition the comments based upon how it adds to the conversation, we could use those to rethink the entire experience of having large-scale conversations. Comments as is may be reading a boiling point of utility at its current scale.

The current classifiers try to guess political affiliation and overall sentiment. While these are useful tools, it would be useful to create classifiers to showcase diversity of thought, or finding perspectives outside the norm for a given issue. Doing so would help provide more balance and nuance in the surfaced demographics. The interface would likely need to be rethought for such information: how can we exhibit textured arguments beyond a simplifying bucket? What would a “fuzzy bucket” look like that reflects its internal disarray?

Buckets themselves are one approach, but future versions should consider the wide variety of user goals and effort available. Some may want a crowd view similar to Anthropomorphs, complete with legible and animated avatars. Many simply want to read a couple of comments before exiting the article. In this case, a production version might consider how to automatically select buckets that are instantly available to read select comments, while still providing summary and exploration tools for the curious. One issue with automatic choice is undue bias: if we allow individuals to preselect the types of demographics they wish to see, we risk creating an information filter bubble. The beauty of the current design is that it presents its demographics per filter in complete, requiring that even if you want to see what Republicans think, you still must see that Democrats or People who mostly discuss sports exist and may be a large part of the population.

Participation itself could be more lightweight than the costly submission of comments. Facebook has already shown how popular a 1-bit *Like* button can be, might a “*Disagree*” button be useful next to a comment? BuzzFeed has employed a small set of tags for comments such as *fail* and *lol* with high participation rates. Participation could also be more heavyweight by letting the user control self-presentation. Defuse was built using a empirically-driven philosophy, but that does not preclude letting users annotate or organize their own data. A user might wish to explain why their interests varied over time, or how individual comments have shaped their world view. Participation should also extend to the filters and classifiers themselves, building tools to crowd source the computational deconstruction of habitus.

As mentioned previously, Defuse suffers from sparsity issues in user participation. If we were able to create composite identities by brining together more data sources of about an individual, we could form a much better data portrait of a their opinions and habits. As so many aspects of our lives are already online and more soon to come, not to mention trends of data linking and

sharing like OAuth and RDF, this is not a far-fetched reality. It is required to go beyond the limits of character that any one context would show.

REACTIONS

The final version of Defuse (*Version 4*) has only been recently released on the web as of this writing. As such, it has not received the kind of attention that Personas enjoyed. While it was evaluated in depth by a panel of domain experts in the next chapter, it also did advance to the semi-finals in the Mozilla+Knight Journalism (MoJo) challenge: Beyond Comment Threads (we withdrew given their required additional lecture and homework during the thesis defense period). The competition looked for new ways of implementing comments on the web to enhance democracy and participation. In the spirit of openness, the challenge publicly listed entries and participants were encouraged to comment on each others submissions (although this rarely happened in practice). These were the two comments left on Defuse's MoJo page:

June 16, 2011, 4:14 p.m. - openiduser318

Great idea! Seems very useful

June 20, 2011, 8:24 a.m. - manus

Really cool idea, and niiice prototype! Still, I wonder whether the benefits of exposing data about an individual commenter outweigh the potential for increasing a reader's bias for/against any given comment by that commenter?

This could definitely be a great tool for visualizing the sociopersonal (is that a word?) landscape of a comment thread, site section, an entire site's userbase, or even for self-analysis. Still, I'm a little afraid that it might enable users to become more prejudiced against comments/ideas that might otherwise be able to stand on their own if these data are exposed for individual users, interesting as they may appear.

SUMMARY

A long-standing notion in Computer Mediated Communications is that the current communicative act is the main way to represent an individual. We see their comment on a particular article, but their 1,038 previous comments are hidden away onto a "profile page" which just lumps them unintelligently in a giant linear list. Defuse is a person-centric commenting interface that uses an individual's collection of digital traces to create a data portrait of them, and through aggregation, the crowds in a given community. It focused on using data from NYTimes.com. BestBuy.com product reviews were also attempted, but were found to be too sparse.

There are many insights that can be empirically discovered using a user's history of comments. Here we judged users using filters motivated by larger sociological frameworks using community-specific statistics. The key in our approach is to find the metrics that are most meaningful and intelligible, while remaining as objective as possible. The first iterations of Defuse relied too much on 1:1 mappings to stay objective, eventually resorting to more potentially subjective abstractions so as to communicate at an appropriate semantic level to the user.

Creating meaningful filters not only yields a powerful top-down view onto the comments themselves, but also a basis to navigate a stream of comments along the dimensions which already fragment society and communities. Defuse can act as both a digital mirror to individual contributors and the community as a whole.

∞. Section review

Four distinct technologies and designs were created in this thesis. They are as described in the previous section: 1) *Is Britney Spears Spam?* (Zinman & Donath, 2007), 2) *Landscape of Words* (Zinman & Fritz, 2010), 3) *Personas* (Donath et al, 2009; Zinman & Fritz, 2010), and 4) *Defuse*.

The work represents this researcher's arc through trying to expose and understand individuals and crowds in social spaces against the backdrop of an explosion of social media. They progress from a metadata-level focus and slowly shift into content analysis and exposure. They each differ in goals, aesthetics, community type, and representation. *Is Britney Spears Spam?* expands on the meta-data analysis by examining the structural-level behavior of users within a space. It shows that these behaviors differ amongst users in prototypical form, and are able to predictively align with a model of a single subjective perspective. *Landscape of Words* begins the journey of content analysis, providing a visualization of an entire community based upon their themes. *Personas* switches from the extreme of the crowd onto the individual. It expands on the NLP techniques of *Landscape of Words* to play with representation of heterogeneous textual characteristics of a name at Internet scale. Finally, *Defuse* uses a hybrid that bridges individual and crowd-level Internet portraits. It segments users into categories inspired by Bourdieu's notion of habitus, providing a method to see the demographics of a community before its comments. Defuse creates a public space that is partitioned by holistic personal characteristics rather than the ephemeral aspects of any one communication act.

5. ABSTRACTION TECHNIQUES

The previous chapter described four unique experiments that all used machine learning in some form. Machine learning offers a useful set of techniques for synthesizing and abstracting details about strangers and crowds. Unsupervised machine learning can find emergent patterns in the data without cultural models to bias the results. Supervised machine learning can be used to classify individuals or crowds by demographic and personality traits. Semi-supervised learning, a hybrid of the two, could allow human users to inject small amounts of their intelligence through annotation, and then let the machine extrapolate the rest of the data. While the use of machine learning sounds great in theory, the reality is much more difficult. Natural language processing still remains difficult, and it is often hard to know what to expect with messy real world data¹⁴, as most published algorithms use cleaner data sources such as the *Brown* corpus which is a compilation of 500 English texts from 1961. The *data modeler* truly has to play and massage their data in a process that is often more art than science. This chapter catalogs this researcher's experience with characterizing social media data, including some methods that did not pan out. It outlines three principal uses for machine learning in assisting online impression formation: 1) summarizing text, 2) finding characteristics of language use, and 3) identifying personality traits and prototypes.

5.1 Summarization

Extracting high-level meaning out of text has been an ongoing challenge in Natural Language Processing for decades (Manning & Schütze, 1999). The existing developments power systems like *Landscape of Words* to automatically cluster and visualize an entire corpus into a common map. To accomplish this task and similar, we need methods to automatically extract meaningful data to allow observers gain a sense of the possibly broad data.

There have been a wide variety of approaches to summarizing large numbers of documents into discrete topics. The three main approaches are classification into domains, vector space models, and keyphrase extraction (Manning & Schütze, 1999). Classification models try to find how well a given set of documents fit into pre-defined categories, typically exploiting highly domain-specific words. For example, it is common to use newsgroups as domain-specific copra to train a classifier (McCallum, 2002) to assign a weighted classification vector or a binary assignment to each document, which can then be aggregated at a corpus level. Popular classifiers include k -Nearest

¹⁴ There are a number of ways in which ordinary writings on the web can be considered messy from the viewpoint of natural language processing. Ubiquitous difficulties include incorrect grammar, misspellings, incorrect use or lack of punctuation, heavy use of symbols such as “ASCII art” and emoticons, conceptual structure through the use of whitespace versus typical sentence boundaries, slang, and mixing or interweaving of different languages.

Neighbors, Maximum Entropy, linear regression, Naïve Bayes, Support Vector Machines (SVM), and Perceptrons (Duda, Hart & Stork, 2000). Vector space models attempt to reduce the dimensionality of documents most typically by representing each document as a vector based upon a word-count, where each possible word in the vocabulary is a dimension and its value for a given document is how often the term appears. This representation is then significantly reduced dimensionally to create latent-space representation. Popular techniques include Singular Value Decomposition (SVD) often employed as a part of Latent Semantic Analysis (LSA) and Principal Components Analysis (PCA), independent component analysis (ICA), non-negative matrix factorization (NMF), factor analysis, Multi-Dimensional Scaling (MDS), Isomap, and autoencoders. Keyphrase extraction methods attempt to find the phrases most representative of a given body of text. While most algorithms have been focused on document-level summarization, some inroads have been made at multi-document summarization. The main approaches include supervised approaches such as ROGUE measure (Lin & Hovy, 2003), and unsupervised approaches graph-based models like TextRank (Mihalcea, 2004) and LexRank (Erkan & Radev, 2004), centroid-based models like MEAD (Radev et al., 2004), and noisy channel models for sentence compression (Knight & Marcu, 2002).

While classification models are especially legible -- their output is a percentage assignment to a human-readable category -- they are not generalizable when the underlying domains are not known. They also hide much of the linguistic structure that belay social proxemics (Trudgill, 2000). Vector space models allow unsupervised learning to occur, whereby we can use the underlying statistics of the raw text unencumbered by biased outside models. The problem is how to find a suitable representation from a vector space model that can be surfaced to the user in an intelligible fashion. Term vectors already remove much linguistic data in the construction of sentences. Automated summarization techniques are less vulnerable to semantic ordering and linguistic issues, as they preserve much of the original wording. However, while some inroads have been made towards natural language generation (Reiter & Dale, 2000), extracting key words or phrases from a document or corpus is not a scalable approach when looking at millions of documents unless those documents are extraordinarily similar.

TOPIC MODELS

Fortunately, there have been recent advancements in vector space models in the form of so called topic models that have an underlying representation that is more easily comprehensible by humans (Blei & Lafferty, 2009). Topic models employ the principal that all documents are written with a select number of *topics* in mind in varying proportions. Each topic is a set of words that belong to a given topic with varying proportions. Ideally each topic is well characterized by a distinct set of high probability words. If a topic has such a strong coherence, then it can be presented to a user by showing the most characteristic words.

Topic models hold great promise in serving to gist a large corpus in an unsupervised manner. They provide a kind of semantic compression of word tokens that lead to models that a) explain the main themes of a corpus or set of documents, b) are predictive of meta-data when built into the model, c) provide high-level document similarity capability without being limited by word overlap, d) word sense disambiguation, and d) can show emerging themes or other bursty properties over time. However, evaluating the results of topic models is a challenging endeavor. What might better explain a corpus to a computer may not match the explanations, concepts, or level of semantics that a human would infer. Limited attempts have been made at assessing the quality of topic models in terms of human gestalt (Chang et al., 2009).

The industry staple topic model has been Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003). While many extensions and variants have been created such as Dirichlet-Multinomial Regression (DMR) (Mimno & McCallum, 2008), the Author-Topic (AT) and Author-Topic-Recipient (ATM) models (McCallum et al., 2007), and Dynamic Topic Models (Blei & Lafferty, 2006), LDA generally performs as well or better for the task of finding the main themes of a corpus. See Table 4.1 for an example output on a corpus derived from Myspace profiles.

LDA is a so-called generative Bayesian model. The generative model captures the act of “generating” each word, document, and corpus in a mathematical distribution, whereby the structure that interrelates these is afforded by Bayesian logic. A corpus D contains M documents, where each document w represents a mixture of topics θ out of K possibilities generated by the Dirichlet prior α . There are N words in a document as generated by a Poisson distribution, and

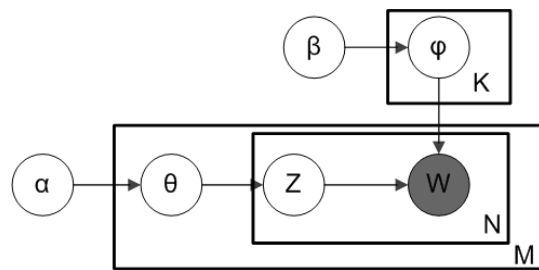


Figure 5.1. Plate notation for LDA. Each plate represents a Bayesian hierarchy with a multiplicity of child nodes as written in the lower right-hand corner.

each word w_n is generated from the word-topic Dirichlet prior β conditioned on a topic z_n . Thus each word is assigned to a given topic z according to the topic-word probability vector ϕ . The plate notation for LDA is shown in Figure 5.1.

Any given set of documents may be fit to the generative model to produce the document-topic prior distribution θ and topic-word distribution ϕ . Techniques exist to then apply these priors onto unseen documents (Griffiths & Steyvers, 2004).

While θ represents the traditional weighted vector of category membership that applies to both classification and latent vector space models, ϕ can more easily be presented to the user than vector space models that do not have an easily comprehensible abstract space, as ϕ is a set of words that belong to distinct clusters of topics.

LATENT DIRICHLET ALLOCATION EXPERIENCES

LDA was attempted in three out of the four experiments because of its successful ability to find interesting clusters. Figure 4.19 shows sample results from *Personas*, which are very impressive given the complete heterogeneity when searching arbitrary names. LDA is not a black box with an on/off switch, it requires finesse and parameterization to produce quality results. We discuss some of the key issues.

When attempting to run LDA on a set of data, there are three variables that must first be specified. The first two, α and β , construct the Dirichlet priors for a given topic or word, respectively. The third is k , the number of topics. α and β are relatively easy to fudge: most LDA implementations recommend the use of $50/k$ for β (I find 0.01 to 0.12 adequate), and α to be 0.1 for tightly specified documents and upwards of 1.3-2.3 for normal social media text. Mallet offers hyper-parameter optimization and asymmetric priors to lessen the need to predict the ideal α and β (McCallum, 2002). Setting and knowing the appropriate k is the trickiest proposition. If k is too low, we do not capture enough of the variation in the data. If k is too high, we create topics that repeat themes or inserts garbage because the model wants to put something in that topic. We cannot rely on automated meta-optimization approaches to find the appropriate k for an arbitrary dataset because we have no decent measures for semantic cohesion of the resulting set. Perplexity (Brown et al., 1992) has been shown not to correlate well enough with human judges (Chang et al., 2009) and in informal practice. Finding the right number of topics is often not just a match with the inherent semantic qualities of the dataset, but also represents a compromise in the user interface. In *Landscape of Words*, we originally set k much higher with excellent results. However, the total number of topics did not fit well onto the landscape because there were too many topics to traverse spaced very close to one another. *Personas* suffered a similar fate. Thus for simplicity on the interface we sacrifice semantic purity by overloading topics. Setting k artificially low will force less probable words and domains to merge with dissimilar topics.

LDA works by looking at every word in its corpus and assigning it to a topic. While this works well for very domain specific words that have high co-occurrence probabilities with semantically similar words, natural language contains many other words that qualify other words and build metaphors, concepts, and narratives (Lakoff, 1987). These words, such as *around*, *time*, *above*, have little or misleading meaning on their own. For example, the sentence “This *time* I mean it!” is about communicating certainty, not time. Yet unless removed, these high probability words will scatter across the topic model and increase error and reduce legibility. In the course of building many topic models, I increasingly became aggressive about adding such words to a central stoplist at the cost of reducing accuracy in some edge cases. Most stoplists are under 50 or 100 words, Tokup has over 800 words in its stoplist. This list has been open sourced along with *Tokup*, a Python-based tokenizer for social media data described in the previous chapter.

While LDA has worked well at a corpus-scale of millions or hundreds of thousands of documents, it has experientially worked terribly for small scale corpora. This is because it must fight Zipf's law, which states that the frequency of any word is inversely proportional to its rank in a frequency table. As most words seldom occur, there needs to be a critical threshold of co-occurrences to tightly form meaningful topics. Otherwise the topic model will haphazardly consist of high-probability words across the topics. While the ideal number is a function of length of document and the natural distribution of semantic tightness in domain-specific words, it has been found that at least 5,000 to 10,000 documents are needed before meaningful topics may arise in social media data.

As LDA is Bayesian, it can be extended with many more signals beyond the bag of words. Future work should examine the role of many socially meaningful signals altering the model such as those found in *Is Britney Spears Spam?*.

5.2 Finding characteristics of language use

How one writes is very telling of their socioeconomic position and more (Bourdieu, 1979; Bonvillain, 1993; Thornton, 1996). Beyond semantic preferences, aspects like grammar, misspellings and word choice function as informative sociolinguistic markers that aid prototype formation (Hudson, 1996). Many of these aspects can be broken down into computationally tractable chunks in ideal conditions, such as tagging parts of speech to determine writing style (Manning & Schütze, 1999). However, when dealing with real world social media data, much of the source material is not grammatically-correct enough to rely on parsers. This author favors finding sociologically meaningful signals that can be either presented in a 1:1 fashion to a human, or to take a more empirical n -gram approach as to rely less on erroneous NLP.

There are many 1:1 signals that meaningfully split audiences to human readers. For example, we can look at the histogram of casing to differentiate PEOPLE WHO WRITE LIKE THIS versus the people who write in all lowercase versus, the people who correctly capitalize I. Such a statistical approach simplifies the need to heuristically determine correct capitalization, which can be expressed in the interface on relative terms. Defuse uses percentiles to bin users into capitalization prototypes. Aside from capitalization, it is easy to compute many other sociolinguistic facets that can be easily interpreted by users: spelling errors, "I" vs "you", length of message, histograms of punctuation kind and frequency, presence of ASCII art, histograms of emoticon kind and frequency, frequency of hyperlinks, and presence of vulgar or slang words.

However, we would also like to find cultural references or unique ways of phrasing ideas to communicate one's habitus and sophistication of thought. A statistical approach favors finding how individuals differentiate compared to the norms and highlighting those raw snippets. We tried two main techniques on our dataset with mixed results: collocations and significant phrases.

They worked poorly. Future work might attempt to use annotated topic models from different corpora to classify subjects sociolinguistically. The topics found in Myspace as shown in Figure 3.1 demonstrate the possibility.

COLLOCATIONS

Collocations are refer to sequences of words that co-occur frequently. For example, “crystal clear” and “nuclear waste” are collocations. If we violate the shared lexical structure the sentence will read as awkward. For example, “glass clear” and “nuclear litter” feel odd even if they retain similar semantic structure. In practice, automated techniques at finding collocations turn up names and other cultural references.

Collocation detection is difficult because we must discern meaningful phrases from semantic constructors such as “by the.” Typically an n-gram model is built of a corpus, filtered for desirable parts-of-speech such as nouns, and hypothesis testing is employed for each n-gram (Manning & Schütze, 1996). For experimental purposes, the BestBuy.com product review dataset was analyzed using LingPipe’s implementation of collocation detection. The results, as shown in Figure 5.2, identified many celebrity, band, and movie names but did not surface any other type of sociolinguistic markers. Similar results could be obtained through tagged corpora such as Freebase.

SIGNIFICANT PHRASES

Finding significant phrases is similar to collocation, except that we expect a longer sequence of words that hopefully demonstrates a character trait. These are phrases that statistically occur at higher than expected probabilities for a given individual. Therefore if someone used the expression “the dc fat cats” in every message they posted, hypothesis testing could surface the expression. This technique could in theory spot astroturfers or other talking points-oriented mouth pieces. Unfortunately in practice this has resulted in a lot of garbage when profiles are overly sparse. Figure 5.3 demonstrates the poor results on BestBuy.com users.

5.3 Identifying personality traits and prototypes

Characteristics from sociolinguistics go a long way in gaining an impression of a stranger. Yet focusing on 1:1 mappings of statistical properties or discovered phrases operates at a lower semantic level than what is often useful for a set of goals. If we desire to prototype strangers into existing semantic units, we need to further abstract the data into either known categories as Defuse or emergent categories with a human interpreter as Personas. This roughly translates into supervised versus unsupervised processes. Both can be successful with social media.

Rilo Kiley	Costa Rica
Shivering Isles	Il Divo
Dire Straits	Battlestar Galactica
Janis Joplin	Rascal Flatts
Ong Bak	Brandi Carlile
Avenged Sevenfold	Consumer Reports
Subspace Emissary	Altec Lansing
Iwo Jima	Cate Blanchett
Tomb Raider	Resident Evil
Tobe Hooper	Best Buy
Wanderlei Silva	Papa Roach
Satoshi Kon	Winona Ryder
Indecent Seduction	Monty Python
Yueh Hua	Guitar Hero
Val Kilmer	Ric Flair
Kung Fu	Milla Jovovich
Eleanor Rigby	Jada Pinkett
Gogol Bordello	Arrested
Uma Thurman	Development
Mass Effect	Puerto Rico
Mortal Kombat	Ashton Kutcher
Geek Squad	Vin Diesel
Los Angeles	Takashi Miike
Notre Dame	Leisure Suit
Amon Amarth	Casino Royale
Dolph Lundgren	Bram Stoker
Gran Turismo	Brad Pitt
Arkham Asylum	Roller Coaster
Modern Warfare	Meryl Streep
Blackie Lawless	Sigourney Weaver
Killswitch Engage	Scarlett Johansson
Carmen Electra	Heath Ledger
Keanu Reeves	Katherine Heigl
Elder Scrolls	Grand Theft
Hayden Christensen	

Figure 5.2. Results from LingPipe’s collocation analysis on BestBuy.com product reviews.

So that prototypes are easily explainable, they should ideally be related to *habitus* or other forms of social fragmentation. For example, the political affiliation filter in Defuse segments users into easily recognizable prototypes Republican or Democrat. Is Britney Spears Spam? might have used a firewall-metaphor, but it is much more difficult to understand the difference between a 3 or 4 in promotional intention than it is to hear about race or education level.

Of course we run the danger of prototypes being unequally weighted in societal bias and thus should be aware of the conjured images. For example, focusing on race and religion can trigger more bias and assumptions than dress style or the make of car. Expressing prototypes around taste can be an elegant solution as they are known to proxy for more divisive labels (Bourdieu, 1979). Surfacing the raw taste space might be tasking for the observer; techniques exist to contextualize individuals within the larger taste fabric (Liu, 2007).

In creating recognizable prototypes, simple classifiers using term frequency and related kernels often work well. A bag of words model can be successful due to the behavior and environmental priors for a given target population. One type might have preferred subjects and vocabulary, whereas another might

likely be rooted in a different socioeconomic and thus educational geometry. Thus certain words will statistically be harbingers of a given class. Defuse implements political affiliation in this way, using Maximum Entropy. As an example procedure, we outline its construction:

A training corpus was created by subsetting the larger NYTimes.com corpus for any message that contained the word *democrat* or *republican* and their variants. That resulted in a drop from over 2 million comments to roughly 115,000. Of those, 5,000 were randomly selected and given to paid

increasing tones of continue overall great sound great 10the went into putting first exit the and motions are perfect but close	has a lot these speakers to listen to music you have to for this price	the plot itself excellent job as worth the watch does an excellent watch if you worth the buy an excellent job is definitely one an alright movie and the acting with some great definitely one of
the battery insert rechargeable battery insert and the wiimote in contact with on the dock contact with the the rechargeable battery the remote and	new ps3 slim the new ps3 this phone system hellions on parade on the go	the multiplayer is the automotion plus
i would reccommend pros and cons	the gutsy geeks	unit on with the unit on on with the with the remote
guns n roses	out how to	
i brought this	cable goes from	cell phone within

Figure 5.3. Significant phrases as determined by LingPipe’s implementation which compares each author on BestBuy.com to the site’s cumulative n -gram model. Individual authors are separated by spacing, highlighting issues of sparsity.

human judges in the US using CrowdFlour. The judges were asked to characterize the author of each post as one of the following choices: *Republican*, *Democrat*, *Independent*, *Other*, and *Don’t Know*. CrowdFlour managed the labeling and quality assurance of the human judges, eventually finishing labeling the data set with 89% inter-rater agreement. Those labeled messages were tokenized with Tokup and used to train various classifiers. Results of those attempts can be seen in Figure 5.4. The results were poor at first (although better than random), so the data was first filtered to three classes by collapsing *Independent*, *Other*, and *Don’t Know* into a single class. The collapsing improved accuracy on the held out test data by roughly 10%, but was still poor. Finally the data was split such that only Republican and Democrat labeled comments were classified, bringing the 10-fold cross validated test accuracy to 67% using Maximum Entropy. Most likely the high training accuracy of Maximum Entropy reflects an over fit model due to the low number of samples (5,000 or less).

Future work should focus on improving classification accuracy by first dramatically increasing its sample size. It is difficult to predict the potential failings of Maximum Entropy without knowing the performance with a large dataset, as often the cardinality of data trumps the choice of algorithm (Halevy, Norvig & Pereir, 2009).

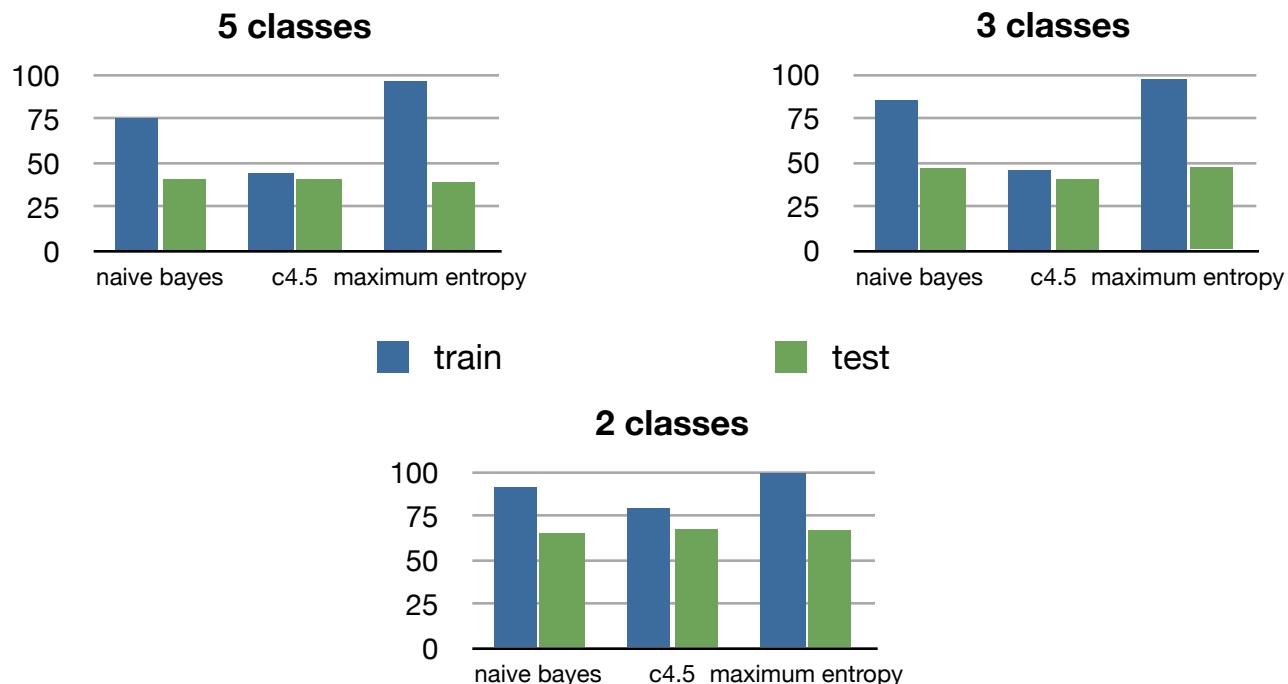


Figure 5.4. Results from training a political affiliation classifier. The training dataset contained labeled messages of up to 5 classes: *Republican*, *Democrat*, *Independent*, *Other*, and *Don't Know*. To improve accuracy, a 3-class version was created by collapsing non-Republican and Democrats into a meta class. Finally the meta class and its associated messages were removed, resulting in a 2-class model which performed the best.

∞. Section review

This chapter has looked at various techniques to synthesize and abstract social media data to improve the perception of strangers online. We have seen that LDA is able to work reasonably well for many corpora to summarize the text in both topic and cultural facets. With a high minimum of input data, LDA works much better for large corpora than small, limiting its utility to create data portraits for subjects with sparse digital footprints. Classifiers of structural behavior can achieve quality results without the need for natural language processing, and as such are a recommended first step. When content analysis is needed, traditional classifiers like Maximum Entropy work reasonably well for character traits such as political affiliation without additional effort. We have also seen other techniques to classify the data less accurately, yielding either garbage or results that were difficult to make useful. Techniques to find collocations find n -grams typically of proper nouns such as celebrities. Looking for significant phrases by way of unlikely n -grams yields poor results. The discussed techniques are but a sampling of NLP literature, however they do represent a wide variety of approaches that were specifically chosen for the task.

Their successes give hope that large-scale, automated, and meaningful abstraction of social media data surrounding individuals and crowds is possible. Future work should examine bringing together simultaneously models of culture, language, community, and society. We need to improve the summarization of subjects with context and surface deeper and more meaningful prototypes about them with higher accuracy to properly realize the vision of better understanding strangers online.

6. REFLECTIONS

This thesis represents the arc of a research journey to expose more of the social life of existing data than is currently practical. It used data portraiture and machine learning to attack the problem of online impression formation: how can we gain a sense of others within a communal context by condensing their past history into a useful representation? This chapter shares some insights gained post-experimentation. First it reflects on the potential to use prototypes in data portraits. Next we discuss what happens when subjects become aware of their observation, before arguing future artists should context individuals and crowds against larger norms. The notion of utility in data portraits is highlighted before concluding with a warning about data type and form.

6.1 Prototyping for interaction

When thinking about how to abstract data about subjects, how we choose which data to aggregate and the methods to do so should be informed by the context of past interactions. Simmel speaks of the need to prototype strangers as a basic function towards knowing how to interact with them. Society has established roles and associated scripts so that basic exchanges can be predictable. We see the bus driver and know the appropriate interaction, and what subtle deviations may constitute a troubling situation. Likewise, online portraits can represent basic prototypes of people necessary to facilitate a social or commercial function. For example, Is Britney Spears Spam? tries to boil down some essences that broadly characterize approaching strangers. Automated prototype abstraction allows a narrative to be followed more quickly or easily when they match the expected or prototypes from an observer.

Bourdieu says that when we perceive, we apply our habitus to place them in the larger social geometry to determine our interaction style. Consequently, we choose to distance or engage based upon levels of social, economic, and cultural capital. Deeper levels of engagement require these pieces to be filled in as they are fundamentally embedded within society. Therefore these geometries should be considered as anchor points in data portraits to push what observer and artist goals they can achieve. While it would be very garish and often in poor taste to outright label people according to these terms, we can proxy these aspects by exposing aesthetics (Bourdieu, 1984). Inroads have been made at computational models of aesthetics that can aid in modeling these spectra (Liu, 2007).

6.2 Behavior awareness

People behave differently in private and with smaller audiences than in public, especially as they will assume different desired perceptions. If we wish to maintain a specific external identity, we

will modify our behavior to best suit that perceived identity regardless of the setting or details of an interaction (Goffman, 1959). Should data portraits become more common, we lose the ability to assume authenticity of the data collected. Instead of communicating for the audience at hand, a different future audience gains the true attention. This tension and its socializing biases have already entered mainstream culture through exposure to social networking data for prospective employees and students (Bell, 2011). It also has the affect of altering non-communicative acts, such as the choice of music during personal consumption. Services like Last.FM publicly broadcast the past tracks played in its users' digital media players such as iTunes. The subject's awareness of their observability collides private preferences with public presentation. This conundrum affects the range of data and perspectives possible in online portraits should they become popular given their very existence will inauthentically modify future data.

6.3 Norms

We gain impressions by judging data against background prior distributions in culture, language, community, and society. In order to choose an appropriate semantic unit, we must make biased decisions when creating a model. Our biases may come from norms expected in the common ground (Clark & Brennan, 1991; Clark, 1996), or they may be created by normalizing data from the community at large. The second is a particularly interesting affordance of data portraits, in that they have the capacity to let observers make more accurate judgements by exposing the empirical and natural patterns in a community. Humans are much better at relative judgements than absolute, and thus it is more fair and useful to give a community comparison basis (Ariely, 2010). This can be achieved in choosing data labels without showing raw values. Defuse uses relative "rare," "often," and "frequent" community norm-determined terminology.

Comparisons against community norms not only help with issues of subjectivity, but also help signpost the rules and demographics of a possibly unfamiliar community, as is the basis for Landscape of Words and Personas. When a user stumbles upon a given web site, they may not know the intended audience and thus the common ground for the participants, or what kinds of users they may expect on the site. For example, a comment that receives 105 recommends on NYTimes.com may seem high, but it may actually be typical for that community to receive hundreds or thousands of recommendations. Trending topics on Twitter helps signpost the types of emergent network conversations versus the more insular uses of the medium, and identifying which tweets in a user's history came from a trend could help establish the intension behind their messages.

6.4 Utility

When online impression formation is presented in terms of observer goals, its solutions will be given utility-driven focus. It is assumed that users know what they want and will discard everything else. This is often not true. Many users cannot or may not be able to articulate a goal, and instead simply want to browse, indiscriminately chat, and absorb information loosely. The popularity of Flipboard speaks to the desire for information as goal-less entertainment. Thinking in depth-based goals rather than breadth-based surfacing of trends also masks potentially key aspects about this world that are important for civilization and liberty (Mills, 1869), promoting the rise of information bubbles (Pariser, 2011). When we abstract individuals we should not only aim to describe them within a narrow description, but instead make sure some more human essence is preserved. When we abstract crowds, we must make sure to leave visibility to demographics that may not be desirable, but are present. This is especially true of those already at the margins of society.

Utility-oriented interactions naturally shift into the more social. This is particularly true of MMORPGs, where social engagement occurs on the side lines of game strategy, sometimes becoming the main focus (Yee, 2007). Providing a more general representation of a person's interests can provide a catalyst for such interactions. While it can be difficult to predict how these contexts may shift, interactive data portraiture affords multiple perspectives onto the dataset and lets the user more elegantly make this shift rather than the developer. Unfortunately serendipity is hard to plan or force onto an interface (Eagle, 2004). Such side effects of a narrow utilitarian intention are better served by preserving more aspects of humanity during abstraction.

6.5 Data types and issues of form

Not all data are created equally. Some data are in forms that are harder for a machine to understand: videos, images, and even text due to natural language processing limitations. When considering condensing an online history, we have to weigh issues of machine perception against importance of any one datum. Some services create supportive structures in the network to sideline machine perception issues, such as Facebook's 1-bit Like button or YouTube's thumbs up/thumbs down. These methods are helpful in determining some elements of rank, but ultimately they do not combat the larger issues of abstraction. Until NLP and computer vision make significant progress, we have to at least acknowledge where data may be missing from the larger portrait. Artist endeavors into video summation can help indicate activity while the larger AI problems are solved (Viégas et al., 2004; Kubat et al., 2007).

The most interesting aspects of an individual may be the form of their communications rather than the content or its perceived likeability. Digesting form into a narrow description is difficult for humans and machines alike. Shakespeare often used iambic pentameter, but mentioning that alone does not do his writing justice. Inroads into modeling linguistic style have yielded limited

categorical results (Newman et al., 2003; Liu, 2007; Mukherjee & Liu, 2010), at some point the raw data must be exposed. Smarter algorithms may choose which data to show based upon what can be inferred, such as structural network behavior, but random samplings may be necessary to achieve a more balanced perspective.

7. EVALUATION

To assess and to gain additional insights into the problem, vision and proposed works of this dissertation, a panel of domain experts was assembled and interviewed. Each were chosen to provide a unique perspective onto these issues. Members include social media researchers danah boyd, Ethan Zuckerman, and Howard Rheingold. boyd, a senior researcher at Microsoft Research specializes in social media practices and how it intersections with society. Zuckerman is a researcher at the Harvard University Berkman Center for Internet and Society, focusing on issues surrounding globalization and the Internet. Boston Globe creative technologist Chris Marstall was interviewed to get the perspective of a site owner, along side Alzheimer Research Forum (Alzforum) creator and Adam Lambert superfan and author June Kinoshita. For a particularly interesting user perspective, Linux kernel developer and anesthesiologist Con Kolivas was chosen for his need as an outsider to virtually establish himself within the Linux community.

Each expert was interviewed in a semi-structured format. They were given a 30-40 minute presentation on the problem, the approach, and the work. Afterwards, they were asked questions in a loose survey attempting to tease out their insights into the importance of the problem and its definition; the existing solutions, their advances and their disadvantages; reactions to the experiments individually, their strengths, and their weaknesses; possible alternative approaches or considerations for future work; and finally thoughts on the overall vision of improving impression formation online through abstraction of digital footprints. This section assembles their quotes and concepts, roughly organized in the narrative above.

7.1. Evaluating the problem

While Rheingold felt that the problem of impression formation online was *“important to individuals and important to the commons,”* others on the panel were hesitant to rank the problem in its importance. Marstall felt it *“does not rank on the top 12 problems of the Internet.”* However, all agreed that the problem does affect people now and will only continue to do so. Kinoshita observed that *“people who dismiss these relationships as superficial just don't know what they're talking about. I didn't use to believe it was possible [for them to be so deep].”* She continued to recognize as a site owner the need for tools to catalyze the discussion; *“I would love to have people be able to connect with each other and share information more easily.”* As a digital organizer of Adam Lapert fans, she shared that *“it is hard to see all the other communities out there... the circles typically grow to a certain size and then fall into themselves. It is hard to know what's going on. If they're out there, how do you connect? It's like another galaxy that you want to send a spaceship to.”*

Others noted the problem is unevenly split between stakeholders, principally between marketers and consumers. *“I'm not thrilled about targeting, but it is important”* Zuckerman admitted. *“Figuring out*

how to do it well, transparent, and mirror-like is important.” Meanwhile, consumers face a different problem and subsequently would require different solutions as *“so much of what we get is from unreliable cues. If you write 10 words I don’t like, or are from aol.com, I may not read your email.”* As such, approaches like Defuse help with *“an incredibly important civic media question.”*

SCOPE OF THE PROBLEM

Not everyone is a site owner, superfan, or large contributor that has troves of data to go through. Similarly, not all aspects of a person are germane. Kolivas emphasized that *“a focused profile is something people will really want. With linux kernel developers you really want to know [about them] in respect to the opinion of their code. The same is true with a motorbike; you don’t care if they’re a member of an illegal motorbike gang as long as they took care of the bike. You don’t need to always have a global view.”* Here Kolivas highlights the importance of goal-driven perception for more utilitarian contexts. This is more difficult to translate in a universal fashion when the goals can be so wide-ranging. However, data portraits can still reflect the compartmentalization of subjects’ lives.

COMPLEXITY OF THE PROBLEM

The problem is complex, especially because different actors exist with different motivation. boyd teases out how this has changed with time: *“Web 1 was all about interacting with strangers, and all about interacting with people different from you, and the possibilities of all this. Web 2 has been fundamentally about people who already know one another. So there’s a very interesting tension between wanting to have access and wanting to see people who are not part of your friend group, and just exist in the world you already know. One of the challenges becomes: what are the implications of those perceptions? I was asked yesterday to talk about why Myspace failed in comparison to Facebook. Myspace allowed you to very easily see people who were from a very different part of the world. It allowed you to judge and critique them in really negative ways. People projected their own values and interpretations onto Myspace as a service, meaning it quickly became ghettoized because people represented different parts of the world. Meanwhile in Facebook, you have to go through so much more effort to see people different from you. So one of the challenges in making visible strangers is to what degree are you going to encourage tolerance or create intolerance. We go out of our ways to build walls so we only see communities like us. When you do see a stranger, you emotionally project the idea onto them that they’re like you. So in regards to the online world, what is the cost to being aware of race?”* The network effects that build communities also reflect individuals and their connections influenced by homophily. As long as members feel they have a place within a larger context, they should be able to navigate the social situations with the same aptitude as in the offline world. Proxying race and other unnecessarily biasing aspects through taste is one abstracting methodology that avoids this situation, as is exposure and easy access to others knowingly different from oneself.

Others recognized the difficulty of assessing people given the large volume one has to sort through. Kolivas notes that *“the problem is a big one, but it’s always been there. Now it’s just online form.*

Previously with colleagues you had to spend time with them, you couldn't just establish a [virtual] profile. Now you're dealing with people more instantaneously online. You do want an online profile. The signal to noise is a lot harder to sift through." The experiments in this thesis sought to make this process easier.

Rheingold recognized that *"it is hard to characterize people in a small number of dimensions. I might trust you enough to send a check on eBay but that does not mean I am going to let you babysit my kids. Different kinds of reputation are not really fungible, but providing multiple dimensions is important I think."* Thus we cannot expect universal prototypes for reputation; each context requires its own heuristics to process the data into a useful form for the goals at hand.

TENSIONS WITH PRIVATE LIVES

Despite possibly large digital footprints, many important characteristics of individuals may not be in publicly accessible digital form, intentionally or otherwise. Despite ease, many may wish simply to remain disconnected or aspects kept more private than others would expect to be public.

Kolivas, for example, has a large digital footprint because of the Linux kernel but keeps his family and anesthesiology profession out of the Internet's grips: *"Lack of profile is helpful... sometimes what's objectively going on is missed by these kinds of lists. There's a certain unnatural aspect to being socialized to such a large community versus you only work with, the people you know, and establish a rapport with patients as you go along. I find it a bit uncomfortable the transition from the small social group into the larger one. Perhaps I'm providing the opposite version as there's no personal data about me online. Some people will be really happy to have their entire lives online, and others won't want that at all. This resonates with my experience in linux kernel."*

WILL IT BECOME A BIGGER PROBLEM?

There was a variety of perspectives on how the problem will change in the mainstream's eyes as we move into the future. Kinoshita acknowledged that *"creating legibility will lead to new possibilities"* and that *"in any situation people will try to understand what's available and adapt to it. It would be nice if I didn't have to adapt and the tools were just better for the problems. These tools will become part of the repertoire."*

It also was universally agreed that this direction of aggregating digital footprints is inevitable whether from malfeasant intentions or net benefit to all. *"Eventually people will judge you for it before they've even met you making personal and professional opinions. A meaningful profile is useful and would be better than people using google alone,"* said Kolivas. He expressed desire for more user control: *"it would be nice to represent yourself in a positive way, like a resume,"* a theme that emerged with all members. The lack of control is one of the most glaring weaknesses of the experiments and approach in this thesis -- it puts much of the power in the hands of the data portraiture artist. To motivate subjects to release their representations to demonstrate positive qualities, erroneous or embarrassing data must have some level of annotation or manipulation possible by the subject for all observers to see.

Zuckerman saw the issue unraveling through centralized solutions. *“I think Facebook is winning, and single identity is going to become a lot more common. I’m already seeing it based on single sign in. At that point we’re collapsing identities. I don’t think it’s unrealistic to expect a future where a few major identity brokers exist.”* He continued, *“it’s a deeply interesting issue for eBay, match.com, Craigslist, anyone bootstrapping. [Is there a] market for it? In the public perception? Not at all. Will it? We’ll get into identity brokers before the need to search [for this information] enters in the public perception. [You] should think about identity brokerage instead.”* By identity brokers Zuckerman is referring to a hypothetical entity that stores all of your data, and works for you on your behalf to issue it only as needed to protect your privacy. This concept has not been fully realized yet beyond the relatively primitive Facebook Connect. It is difficult to imagine how such services could be achieved given the following difficulties they would need possibly automated methods to determine the minimum data appropriate for the task for each negotiation. Requiring the user to embed sophisticated privacy controls has been shown to be disastrous in practice on Facebook.

Meanwhile boyd was convinced the battle would play out more by centralized powerful companies involved in marketing. *“It’s going to be a gnarly battle,”* she said. *“Whether trying to get regulation through law or... for the foreseeable future its going to be a battle. [You are] trying to have a conversation about social norms, but the main powers are going to be the marketing ones. But they’ll battle it over the norms.”*

7.2 Evaluating the experiments

This section summarizes the panel’s insights and commentary on each experiment individually.

IS BRITNEY SPEARS SPAM?

While this project was not as relevant to most panel members as the other projects, some deeper level of prototyping was seen to be desirable. Kolivas wanted to know more about the social habits of people, such as *“if someone is likely to keep friends, or upset people.”* Kinoshita wanted to know more about the promotional people, *“more about their relevancy in regards to what they’re promoting. They may be ok.. there isn’t enough resolution there.”* In her work, Kinoshita would prefer prototyping to help her understand *“who to follow in terms of those who have an amplifying effect. In that if I tweet something out, they’ll retweet it.”* Klout and other commercial ventures are attempting to do just that.

LANDSCAPE OF WORDS

The visualization was seen as *“very cool,”* but the weakness of navigation was emphasized. Kinoshita wanted to *“find the users that are most expert on a given topic, and zoom in.”* Meanwhile, Zuckerman found it *“hard to know how to navigate. If I was trying to read this in terms of a heat map, it would be helpful to have in clusters in a way that’s not put in this fashion. To parse this visually, I need to learn how to understand 16 reference points. I need to figure out what the labels are. It might be helpful to conceptually cluster better. A sense of directionality would be helpful.”* He expanded that the desire to empirically

visualize the community leads to usability problems. Once you have the statistical base created, *“maybe now we should modify things more conceptually well... empiricism and legibility have some tension.”* While the visualization itself might not be appropriate for future tasks, the abstraction methodology was shown to be quite successful.

PERSONAS

All panel experts found Personas compelling in repeatedly calling it *“intriguing”* and *“provocative.”* boyd found *“the conceptual work behind personas is fascinating.”* However, much like the reactions in the blogosphere there was much tension between the artistic implications and the utility of it. Kinoshita wished that *“when you look at what comes out, [you can] look at why and what it's accounting for. If it surprises you, you want to know more about why. As a tool, you'll want to know what to do with this information. it would be nice to see per nugget why it's there, especially if its been posted all over the place.”*

Most noted the appeal of Personas was rooted in narcissism. Zuckerman explained that the *“first thing anyone wants to do is look at themselves and determine 'how good is this a picture of me'.”* He also found it to be *“a provocation in terms of two things 1) what information is [already] out there, and 2) do I like this allocation, and [if not,] who's fault is it? Is it [Personas'] fault for doing the algorithm that says what it does, or is it my fault for what data is being portrayed?”*

Zuckerman also spoke of an interesting parallel with the *“quantified self”* movement, where participants go to great length to log various seemingly mundane aspects of their lives such as heart rate, sleeping patterns, sexual activity, diet, etc., and eventually visualize this information to improve themselves or share it with others. When thinking about personal representation and the types of data that may exist, *“the relevant aspect is that much of this data is for personal use, and we don't have good models of what it means to share it or what to expect as normal.”*

Despite the artistic representation of authoritativeness, Kolivas had concerns about the larger ramifications of such portraits. *“When it's hard data and it's obvious -- you talk 10 min a day to your mom -- it's easy to categorize. Unless you say, 'this is a pure objective measure, or this is a pure conjecture of the data,' it's too subjective. Or say it's 70% predictable, and give how much weight to the user should put on it.”* Thus he pleaded for artists to emboss their work with machine learning accuracies and data portraiture subjectivities so that observers can properly gauge the confidence in their impressions. The difficulty remains in not overwhelming the user or relying on aspects of statistics that are not well understood across society. For example, despite the sophistication of Wall Street traders, stocks often go down when quarterly earnings are less than analyst figures even if they fall within the margin of error.

DEFUSE

The interface itself was well liked by all experts. boyd noted “[the design] is very strong. It’s clear that [Defuse] feeds [off Personas].” There were some objections to issues surrounding legends and color mappings, but their critique centered on the earlier incomplete versions that had not addressed that issue yet.

The utility of Defuse was seen to be wide ranging. Kinoshita: *“A lot of people have that kind of curiosity about the people commenting, especially with high numbers of responses. As a journalist, you want to be able to characterize the conversation. For anyone else, they’ll likely find it entertaining.”* Zuckerman thought that Defuse had strong and important civic media consequences, as did Rheingold. Rheingold wanted to migrate it to Google+: *“There are all these people who add you to their circles; who are these people? I know they’re interested in me, should I be interested in them?”*

The filters themselves were subject so scrutiny and their results a surprise. Zuckerman found the reading level filter *“fascinating,”* wondering *“what is the readability of comments as a whole? Not a lot of people write at a high level.”* It was suggested that the article itself could be used as a reference point.” He also found a *“higher number of frequent commenters than i thought. I thought it was more drive-by commenting.”*

There were also concerns of information bubbles should users be able to set default preferences to only show specific demographics. Zuckerman: *“When does it become not just a way of seeing the crowd but filtering them. [The option to] only show me people who have X and Y is likely problematic.”* The approach taken with Defuse is to always show all demographics even if the observers instantly focus on the same few. Observers may wish to ignore certain sections of society, but the interface does not let them forget that they do exist.

Similarly, Zuckerman also wondered about these filters being used to fracture the intended audience, and the consequences thereof. *“This is an incredible level of detail for someone who is just glancing at comments. I want to see where I come in. Where did I fit in relation to all of this?”* He worried about being over exposed to those who recommend one’s comments, in that if *“I know that you are part of group X, then I might just want to respond to those who [would] recommend [me].”* Zuckerman also contended the upside in that *“maybe understanding who [recommends you] gives you a mirror to see where you fit within the community.”*

Most members agreed that the exposure to various demographics was helpful. Zuckerman found it useful to say *“here is the diversity on these different axes. Here’s all the variation and richness.”* But he also warned that *“there is a normative piece of this, that says: 20% of comments are written by assholes as determined by [this filter that detects profanity], etc.”*

Rheingold wanted to see an expansion of the filters to better prototype the character of the subject: *“It would be interesting to know when people are disputatious -- someone who is always disputing,*

always arguing, always debunking.” Pushing the concept further, he wanted to know more about the changing social dynamics surrounding a subject: *“Does this person play a role? Is that role always the same or similar?”* Machine-identification of arguments is likely easier than assessing role. Structural-level features can facilitate basic role identification (Gleave et al., 2009), but to identify deeper aspects of social roles such as *“queen bees and wannabees”* (Wiseman, 2002) is likely an AI-complete problem.

Kolivas was concerned about comments being a principal data source to judge individuals. *“Gathering a profile based on comments may not be accurate because there are certain kinds of people who generally post in large quantities in 2forums, certain personalities that post on forms, so you form a very biased list of comments and passing judgement on people commenting on a forum regardless of how well their comments are received is not actually getting an accurate representation of an issue. If you're trying to get a feel for what people are saying about it, if you're based on a distilled down profile of commenters you'll get a biased view. That's a danger I feel. Not that you're getting the wrong profile for that person, although there is that danger, but the problem is you're forming an impression based upon the sort of person who is more likely to comment. It becomes a political issue, and if you're a politician going online seeing what people are saying about this issue, then you get the view of whatever those people who login to comment on that particular issue.”* But he found the ability to filter based upon those who infrequently comment to be useful: *“When I log into forums, I see the same names over and over again, and these new people might have something interesting to say. But of course they get shut down in flame by the people who are familiar with commenting and know how to play the game of counterarguments until the other person gives up. When that happens it's taken as a sign that you won, but that doesn't mean you were necessarily correct in the first place. So there's an art form to commenting on forums. If you place just as much weight on the people who are 100x less frequent, you'll get a better overall profile of what's going on.”*

Despite concerns of how people game forums, Kolivas would like to see Defuse move beyond NYTimes.com onto the Linux Kernel Mailing List: *“There is little in the way of filtering those 500 messages per day, and the more you have to sift through the worse. It would be useful to know what are people working on, and what are they commenting on.”* He also saw the utility in bringing status and expertise into reputations and filters: *“having a vague clue in knowing [commenters on a health form] are a professional and there is hard evidence based on the online data that they actually work professionally in a given field, and thus automatically will have a higher rating as result as that or be flagged as professional, that would be worth knowing so you if you asked a question 'my wife is in pain, what do i do about it?' They can see through 2,000 comments and say these guys are professionals, at least I'll read their comments first.”*

OVERALL

The panel was overall very positive about the experiments, despite their critical viewpoints. Zuckerman found it to be *“really fascinating stuff. [I] like the questions being asked. [I] like where it's been taken. I like the design sensibility a great deal. It's both clean & bright without feeling intrusive. I want to touch things.”* He also wondered how one's data portraiture might be taken outside of its original

context: *“This is such high cost data to produce, people are going to want something more lightweight as well. There may be spaces in which case it is helpful; you've made the concept of your strong personality in commenting as a marketable tool outside of it. For people to apply for jobs by demonstrating how good they are at World of Warcraft is very interesting. Perhaps someone is working dead end jobs, but [they can show that they are] a very capable leader. It is interesting to think about what's useful and when it isn't.”* These are indeed strong possibilities that allow everyday digital footprints to gain value in their aggregation.

boyd *“strongly [valued] the experimentation that's been done”* but noted that she *“can't evaluate it without a population. It's a powerful set of experiments, interrogating through intelligent design. The design is thoughtful and provocative, and clearly grapples with these issues. Do they help? Of course. Does more need to be done? Of course. Nobody is ever done.”* She was looking forward to see after deployment to the masses *“how people evolve with it. People will experiments in unexpected ways. While that's one measure of success, as a designer your job is not to say what concepts are most important but understand what they're trying to do and meet their needs.”* This author agrees, which is why Personas was a successful venture: it was used by millions of people. Data portraiture needs to achieve a higher critical mass to fully tease out the requirements for it to survive, which include accuracy, ability to meet goals, simplicity, and how much it allows subjects to save face. In performing these types of experiments, boyd noted *“the outcomes may not always be positive. But that doesn't mean it isn't worthwhile to explore.”*

7.3 Existing solutions

When asked about existing solutions to the problem, experts cited search engines, LinkedIn, Rapportive and its competitors, Lexus Nexus, classmates.com, asking mutual friends, self-monitoring practices within the community (in case one violates face), and simply *“taking people at face value or not depending on context.”* They were found to address many problems when there are greater consequences, independent of effectiveness. Kinoshita wondered *“how reliable these reputations on eBay are, especially for the more expensive items. Hard to know if people have multiple identities, or are in cahoots. At some level the system is working, but it could be improved with better transparency and ways of assessing people.”*

However, for less seemingly important issues such as friend requests, the existing solutions are insufficient. Kinoshita finds managing Adam Lambert's page on Facebook difficult, asking of incoming friend requests, *“Are they an actual person? Spammer? Every once in a while there's porn or something I don't have the time to go and figure out who they are,”* and as a site owner, *“I'd like to better assess their spamminess. Show me people asking to be friends, and show me a set of parameters to understand who they are to separate out the spammers.”* Is Britney Spears Spam? was an experiment with machine learning, but its model could easily end up in future social networking interfaces, only showing requests from those that meet minimum standards of practice. However, she did not find daily instances of the problem to be overly drastic: *“Typically you take info with a grain of salt unless they write about something that could change your life.”*

Zuckerman postulated that while the data available online may be skewed, it still is helpful: *“One of the reasons I use Rapportive [is] when I get an unsolicited email I get enough data so I can make some interesting conclusions about them. It’s a profile people are not aware they’ve put together, because it’s accumulated through their Google profile, Facebook, and Twitter profile. These are interesting aspects of a person that will give you an impression. It may or may not be accurate, but it’s a contextualization like what a lot of us do in real life. This is a model that’s close to a best practice, and I think making it easier is not a bad thing. But I do think it has all sorts of fascinating consequences when it gets mass adoption.”* Rapportive is useful for dynamic networking individuals such as Zuckerman and businessmen, but their popularity for the ordinary person remains to be seen. Rapportive and similar tools still retain the same paradigm of reverse chronological lists of actions. More sophisticated data portraits will have utility beyond just emails, spreading the concept to possibly everywhere one encounters strangers online.

ISSUES OF POWER

A repeated theme in the interviews was that of power in these tools and what the market has already done in these spaces. *“When capital wants to pay attention to it, they will find all kinds of ways to find info about people”* boyd notes. She warned that designers must be aware of *“when is this about reinforcing power and when is it about reinforcing individuals? Anything that is about surveillance will tend to reinforce existing power structures even if that is not the intension. When am I an employer and when am I a friend?”* Indeed employers will have different goals for a subject’s data than a member of a dating site, and given their resources they may have an edge in the creation of those tools. However, this author contends that there are more ordinary people than employers, and as such, data portraits will have a strong future market with a consumer-focus.

LACK OF PERSPECTIVE IN EXISTING TOOLS

Another emergent theme was the lack of broader perspective in current systems. Kolivas: *“We don’t have a reasonable summary of anything that gives you an overall view. I think these tools show a lot more info, existing stuff is mostly just a flag. You know how many lines of code they put in, or what they have liked, but it doesn’t tell you how to ‘read about it.’ What are the other interesting aspects of a person? The ability to expand is a strength of these projects rather than trying to just see a super simple statistic.”* He further contended that *“the more this sort of data is available, the more useful the types of communication will come. Forums become more useful. The comments on newspapers [will become] more useful. Oor they might, they’re still quite simple. Right now on forums you have the ability to read previous posts, which is simple but I find incredibly useful. If i knew something more about them, then I could better determine if their comment is useful.”*

7.4 Evaluating the vision

The panel was excited but wary about the future possibilities of online impression formation. Kinoshita: *“It is worth while. The quality of what can go online is incredibly valuable. If there are things that*

make it easier, then we can give more people confidence to counter all these scare stories about what happens online.” Zuckerman found the existing experiments and vision refreshing given past takes on the data: *“You as a user have a lot of situations where you trying to figure information about others online and you don't have that info and need to go out and collect it. But there's another set of questions where we know your IP address, and we can hash it with a lot of things. This seems to be from another angle. I like it.”*

ALTERNATIVE APPROACHES

Zuckerman wondered about the utility of online information, especially in terms of credibility both culturally and in terms of data source: *“Past behavior isn't necessarily a good indicator when sparse. You might be able to be more expertise based, and that would play out differently in different places. In West Africa, people say 'Who are you to speak on this?' Here you're saying saying what I care about is what you've done. Theoretically there are other ways to do this. For example you could get credit score and append it with what little data exists and label them as reputable or not. There are any number of ways to overlay a type of reputation on top of something else. This is data you have access to. You're overlaying reputation, writing level, political argument. You can imagine someone's IQ score, SAT scores, terminal degrees, hometown, credit scores, you could do a personality type score, etc.”* This is an excellent point: we need to balance what is accessible and desirable to externally present with statistically significant but private information. He suggested the utility in *“thinking like marketer”* in that *“race, gender, age, and ZIP code gives me a pretty good idea of who you are. What are some of these proxies that exist in the data already?”*

boyd implored the consideration of social science fields outside of sociology. Namely, she suggested that the works performed mostly have been in terms of modern day economics, in that they can be framed in terms of *“modeling and tradeoffs,”* and that economics would help with *“trying to understand [the] decision making structures when you can model something very large. Economics understands how to do that kind of model.”*

DIGITAL MIRRORS

Experts were less coherent in the effects of data portraiture as a digital mirror, which occurs when the subject and observer are the same. boyd: *“It could go both ways. It's an interesting research question. There is great sociological work on how if someone in your family is gay, your level of tolerance of LGBT populations is much higher. It's not clear to me the same is going to operate online. That's why I see a lot of question marks rather than answers.”* When pushed on the possibility of persistence to alter future behavior, she stressed that *“people do not project into the future. They do not think about the future in terms of consequences. The population who thinks about future consequences is a narrow portion of the most privileged in society. They are the ones that think about future consequences; the vast majority of people think about what they have to gain. If all your friends are neo-Nazis, you're going to put that swastika on your back. You're not going to care what people outside think of you. Part of it is that you're going to prioritize the values of the community you most care about even if it's a very small community with respect to the world at large. Thus it's an interesting*

research question to ask ‘as people become more aware of one another as strangers rather than intimates, what are the social consequences?’” Indeed this is a potential roadblock for those who might be embarrassed by their past, which will be undermined by the emergent utility when presented to the world for those who come out on top. Examining the level of control and annotation ability subjects have in their portraits can help minimize the weaknesses of maturing and the human condition.

Zuckerman pulls in the perspective of the quantified self movement: *“it is interesting to see the insights that come from people tracking themselves. How would you qualify what you see and hear. What happens when you qualify media? Anytime you’re logging, you’re creating a data portrait of yourself. Here’s my pulse, here’s my sleep pattern. By logging it it changes your behavior. You still alter [yourself] to preserve what you want the data to show. It’s getting easier to build mirrors. How they are working in qualified self or other ways from Personas, [which is] a funhouse mirror. People want to know about themselves.”*

BIASES IN ABSTRACTION

Many were concerned about the biases that occur in abstraction. boyd noted that every abstraction *“has costs,”* and the designer *“must acknowledge the bias in abstraction choices.”* For example, boyd notes that *“putting political affiliation in there was a choice that projects what the users should care about, even if they don’t. They will read it as something they should care about. When we put these categories, the level of coarseness matters. For people in Boston, they may distinguish between Roman Catholic versus Italian Catholic, let alone Muslim or Hindu. What about tea party versus Reagan Republican?”* As such, *“before you even get into the content about the person, the interface signals what’s important.”* The choice to use the language of data portraiture highlights the acknowledgement in bias. It is the artist’s responsibility to assess the balance between their world view and process against those of the subject and observer. Future alternatives to the current model of data portraiture should examine the role of observers to influence prototype mining and presentation of subjects.

TENSIONS IN REVEALING INFORMATION

Issues of how much information should be revealed, as discussed earlier in this chapter, were particularly sensitive for boyd: *“There are situations where the right amount of information gets you where you want to be, and situations where less information gets you where you want. More information can actually stymie your ability to make a decision. Because information about people is rare and inaccessible in traditional situations, we have a deep desire to get as much as possible. And we always think more is better and will always let us make better judgements. Take it to an online dating context: it’s common for people to want to find every piece of information about a person with the assumption that that more info will allow them to draw a more accurate conclusion about the person. However, what becomes clear is that people are good at projecting any information they want into the known info even if it contradicts. How are people using the information to create mental models? And does that mental model give them more comfort? More levels of security? Or does it actually destabilize their own mental models from being able to cope.”* She gave the example of her own blog, when during the first

six to nine months she filled it with details of *“what it felt like to be raped. After many long conversations with a series of people, I decided to remove those blog posts. People ran across my blog, wondering who is this girl, would go to the beginning and immediately get uncomfortable in a context in a way they were not prepared and were thus shocked. So it wasn’t a conversation and thus became a boundary.”* Thus the question becomes what’s the tension between too much, too little, and the right amount of information for every possible situation? The online world complicates this as in the physical world *“when you’re learning information about someone else, you can learn it in the context of them telling it to you, [versus here] you can learn it flat.”* It becomes unnatural to *“learn it flat,”* as *“people give different stories for different people in trying to contextualize it [for the audience]. The process of giving it is to contextualize it for the social dynamics. What are the social processes that make giving information more powerful?”* Part of this problem plays out in the lack of common ground for future observers. But the lack of awareness, dialogue, and influence of subjects can limit the roles that data portraiture can play. If we are automating the abstraction of a stranger’s accessible data, we cannot expect to approximate the same impression that would occur through a deep relationship. We must settle for the best outcome that serves to everyone’s advantage.

7.5 Thoughts for the future

There were many issues brought up that the experts felt would need to be addressed in the future, depending on the context. The themes mostly revolved around users having more control over how they are shown to other people, what aspects of themselves are being visualized and how much weight each data point is given, and how users could take more control on what kinds of filters are being viewed.

MINED BEHAVIOR

The behaviors of people in a social space are very important, and can be difficult to assess without being a deeply involved human participant. Kinoshita noticed how important gossip and dynamics were to her communities: *“Look at what people say about other people. That’s the village gossip approach. If someone repeatedly behaves in a certain way online, it might be nice to know. It would be nice to look more at social dynamics. It goes up a couple levels... [data] mine the gossip and know who is a liar, who says what about who... it’s tricky.”* Unfortunately identifying where gossip is a very difficult natural language processing, let alone taking the next step of inferring social dynamics and personality types from the rendered gossip.

Zuckerman was concerned about *“what behaviors give you the kinds of data you would want. It’s about what is out there and where it’s coming from. People are going for easily accessible data, [this thesis] is going for less accessible sources. All this stuff we’re measuring is performative.”* Low hanging fruit will always begin any venture for those that are risk adverse. However, as data portraiture and the techniques to build them mature, those innovations will cross-pollinate enough to break new boundaries in utility.

CONTROL & PRIVACY

All experts mentioned issues of control and privacy, particularly in concern with how people are able to self-portray themselves in lieu of the data available. Zuckerman: *“Everyone wants control of how they are presented. Then people want to tweak that all the time. [This thesis is] implying a profile that has a high cost to alter. The missing idea is control. How do individuals see themselves and how can they affect that.”*

He elaborated: *“large scale data is hard to create. It may be me, and I may be ashamed of it, or maybe just aspects of it. That question of representation is tricky and hard. For persistent portraits, you need to think more about control. People are sensitive to enough aspects of their existence to want to camouflage parts.”* The issue of control is a crucial part of any future work. While data portraiture places most responsibility on the artist, mass adoption will require all parties to have a larger say and a method to save face.

Kolivas wanted to *“be able to opt out of data that you won't want. Finding out details that would otherwise be hard for yourself is a problem. Sharing of data is going to be difficult -- we need to say this is professional profile, my friends and family profile, my hobby profile. Then people understand how they are being represented. When there is cross over you're revealing different aspects of yourself that you may not want to. [For example,] I'm a heretic in the Linux kernel community. I'm not a heretic in the rest of my life.”* Rheingold felt that such commentary is a means to achieve the problem as *“it gives you another window onto that person. It's the Goffman giving off business. You're concentrating on giving off but maybe you should concentrate on letting people give.”*

UNIT OF ANALYSIS

boyd was concerned that the individual is often the unit of analysis in some of these works: *“in doing this work, you choose what feels right to you. But you need to constantly remind yourselves that there is a world different from yours. The individual is not the unit of analysis for most of the globe. It's family and tribes. That is exceptionally true of India & China, the largest countries in the world. Design decisions couched in the idea of understanding what the design tradeoffs are different from design [decisions] couched in the idea of what feels right.”* This excellent point reminds us of the need to expand or recognize our cultural expectations when summarizing humans that also live within larger cultural contexts.

OFFLINE/ONLINE SPLITS

It was noted by most experts that offline identity is not the same as online, and there are potential liabilities for assuming more of what knowledge is digital than should be. Particularly in the open source world of mostly virtual collaboration, *“there is something about text-mode that allows you to portray a completely different persona”* says Kolivas. *“The offline person isn't necessarily the same as the online.”*

Similarly, there tend to be large biases in the types of jobs and demographics that have strong online representation. boyd warned, *“the social implications of these systems is to be aware of the people you may be excluding in your design tradeoffs.”* The assumption with this work is that the increasing use of social media and mobile devices will help overcome sparsity issues for those who wish for

exposure. Annotation of data by subjects can help aid observers to understand bias in a given portrait.

COMMUNITY-DRIVEN FILTERS

Most experts wanted the ability to have more user control rather than designer control over the representations and filters. In particular Kolivas “*wondered what the end users might want. Not sure these [filters and representations] are always what's interesting about the people. Would be curious to know the range of things that people would want in a forum or social network site.*” Kinoshita wanted to “*be able to fine tune the different ways i want to slice up the community*” as “*every day I could think of a different way to do it.*”

Zuckerman, echoed by boyd, raised the issue that it is impossible otherwise to entirely predict what users will want considering the culture and its will adapt accordingly: “*There are worthwhile open questions on what we actually want to know about a person. It's about the context. It is difficult to know the balance. It'll be interesting to see over time if you give people forty things to figure out, which ones do they actually drill down onto? We'll only find out through iteration and deployment.*”

∞. Section review

The chapter summarized an evaluation of the problem of online impression formation and the solutions described in this thesis by a panel of domain experts. While the issue of online impression formation was recognized as problematic, there was less certainty about its global importance given the rest of the problems of the online world. The designs were seen as relatively successful in the whole, with each panel member drawn to a particular design in relevancy to their own online problems. Various issues of power, subjectivity in abstraction, bias, representation, datatype, control and privacy were raised.

8. CONCLUSION

Every day, millions of people encounter strangers online. We read their medical advice, buy their products, and ask them out on dates. Yet our views of these strangers are very restricted; we see only individual communication acts rather than the person(s) as a whole. Out of this predicament arises an opportunity for the designer: to use technology and data mining to improve our impressions of others online by visualizing their archived digital footprints.

Today our capacities to solve the problem of online impression formation are limited. Despite the impressive rise of CMC in the last decade planet-wide, the possibilities of a networked world and its ramifications are still being revealed. The shared cultural framework for CMC is still in its infancy. Meanwhile, machines cannot yet understand human data as a human would. Hard Artificial Intelligence has not been solved, and with it difficulty in natural language processing.

None the less, progress is being made. In addition to advances in NLP, storage, cloud computing, processing power, and the financial cost of these resources, more significant opportunities are being created through the sheer amount of human-generated data to expose the very nature of human beings (Halevy, Norvig & Pereir, 2009). The continuing deep integration of Facebook and Twitter into everyday society is pushing the cultural expectation of a mediated social life towards into a new normalcy. The richness and volume of these new digital channels is unprecedented. With each new normal comes the affordance to advance and mediate more. Culture may need time to understand and integrate new channels (Walther, 1992), but soon the arc of social networking will be so culturally understood that we can move past a focus on the people we already know and begin to chart the territory of those we do not.

Our drive to know others online more deeply is already evident in current systems. Massively Multiplayer Online Role-Playing Games require tactics to see the community as a whole (e.g. leaderboards) and judge individuals (e.g. gaming stats). Its rich worlds such as World of Warcraft have shown previously unimagined levels of socioemotional content (Peña & Hancock, 2006); these cultural innovations have been theorized by the Social Information Processing model, which holds that while online relationships take longer to establish, “*CMC can supersede levels of affection and emotion of FtF interaction*” (Walther, 1992). Similar examples abound. Thus eBay relies on “gossip” about the reputation of individuals. The dating site Zin.gl uses self-presentation information on Facebook data to algorithmically match potential dating candidates. DJs peacock their community dedication with their earned avatars on Turntable.fm, mirroring the activities of those in Second Life.

Exciting though as these approaches are, they represent mere pieces of the greater puzzle. They are first-order solves for what “visibility” means in a given community. None has an outward

focus beyond its data silo. None of the designed solutions tackles what it means to be human and to see other people in their fullest dimensions. None lets us see the overall structure as we might survey a parade down the street. We are at a communications bottleneck unless we can better apprehension and resolve the challenges raised by online impression formation.

Online impression formation based upon the useful digestion of a user's history is riddled with difficulties and subtleties. We must find ways to account for the original context and common ground of each communication act; to model cultural and community contexts; to usefully abstract and remove data so as to communicate efficiently with observers; to mix and represent heterogeneous data types equally; to remain conscious of the subjectivity in representations; and to retain the serendipitous aspects of humanity in spite of their unapparent utility. These are all difficult issues which designers of CMC must consider in their work.

Through experimentation this thesis has addressed each of these difficult issues in four original designs. Each experiment externalizes previously hidden social fabrics of an existing online world with a unique perspective on the problem of and opportunities offered by online impression formation through data portraiture. Data portraiture, the depiction of people through a rendering of their words and actions rather than their physical bodies (Donath et al., 2010), recognizes the inherent biases of an artist in choosing which aspects of a person to reveal and in how it will then be interpreted by observers. The experiments in this thesis attempt to advance online impression formation with specific goals of the observer in mind without giving power to the subjects to manipulate their portraits. They collectively foster the perception of what is possible with data portraiture, informing future designs that may consider participation by the subject in their depictions.

The first experiment, *Is Britney Spears Spam?*, addresses the challenges of labeling spam in a social networking context. It offers a model to computationally prototype strangers at first contact according to their perceived social and promotional intentions. The model demonstrates that varying behavioral strategies can be detected at a structural level without content analysis. Next, *Landscape of Words* shifts perspectives towards crowds by aggregating and visualizing the content of large online publics. *Topic models* -- the algorithm genre employed for the task -- provide a less biased representation of conversation topics and sociolinguistic markers, and is thus useful as the basis for comparisons of individuals and collectives. Following on the use of topic models, *Personas* scours the web looking for information characterizing a desired name. As a data portrait, it exposes the underlying process of data mining that is normally hidden in a supposedly authoritative presentation. It calls into question a future where our online selves are more important than our offline reputation. Finally, *Defuse* brings together representations of crowds and their constituents through content and structural analysis of user comments. Developed

using NYTimes.com data, Defuse uses sociologically grounded metrics to navigate and understand strangers.

In sum, these unique designs offer a vision in which online communication and behavior break out of obscurity and transience acquiring some of the resonance of in-person interactions and enabling participants and observers alike to apprehend users in their fuller human dimensions. Many technical challenges in computationally processing natural language and human behavior were overcome in the process by applying existing techniques in new ways. A world without strangers may be a utopian ideal for some, but the aggregation of digital footprints promises a deeper engagement with unknown persons and collectives—an engagement that stands to benefit us in every sphere of our social lives.

9. BIBLIOGRAPHY

1. Adamic, L. A., and Glance, N. (2005) The political blogosphere and the 2004 US election: divided they blog. *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pp. 36-43.
2. Albrecht, K. (1980). *Brain Power: Learn to improve your thinking skills*. Simon & Schuster.
3. Ariely, D. (2010). *Predictably Irrational*. Harper Perennial.
4. Barbaro, M., and Zeller, T. (2006). A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*, August 9, 2008.
5. Barlow, T., and Neville, P. (2001). A comparison of 2-D visualizations of hierarchies. *Information Visualization, 2001. INFOVIS 2001*, pp.131-138.
6. Bauer, M. (2000) *Classical Content Analysis: a Review*. In Bauer, M. and Gaskell, G. (eds.) *Qualitative Researching with Text, Image and Sound*, Sage, London.
7. Baym, N. (1993) Interpreting soap operas and creating community: Inside a computer-mediated fan culture. *Journal of Folklore Research*, vol. 30, no. 2/3, May 1993, pp. 143-176.
8. Bell, M. (2011). More employers using firms that check applicants' social media history. *Washington Post*, July 15 2011.
9. Bergstrom, T., and Karahalios, K. (2007) Conversation votes: enabling anonymous cues. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems: CHI'07*, San Jose, CA, pp. 2279-2284.
10. Blei, D., and Lafferty, J. (2006) Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.
11. Blei, D., and Lafferty, J. (2009) Topic Models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, in press
12. Blei, D., and McAuliffe, J. (2007) Supervised topic models. In *Advances in Neural Information Processing Systems 21*, 2007.
13. Blei, D. M., Ng, A. Y, and Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993-1022.
14. Bonvillian, N. (1993). *Language, Culture and Communication*. Prentice Hall, Englewood Cliffs, NJ.
15. Borges, J. L. (1941). *The Total Library: Non-Fiction 1922-1986*. Translated by Eliot Weinberger. Allen Lane, The Penguin Press, London, pp. 214-216.
16. Bourdieu, P. (1977). *Outline of a Theory of Practice*, Cambridge University Press, Cambridge, UK.
17. Bourdieu, P. (1979). *Distinction: A Social Critique of the Judgment of Taste*. Translated by Richard Nice for Routledge & Kegan Paul Ltd, Cambridge, MA, 1984.
18. boyd, d. (2006). Friends, Friendsters, and MySpace Top 8: Writing Community Into Being on Social Network Sites. *First Monday*, 11 (12).
19. boyd, d. (2007). *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media Volume (ed. David Buckingham). Cambridge, MA: MIT Press.

20. boyd, d. (Forthcoming). White Flight in Networked Publics? How Race and Class Shaped American Teen Engagement with MySpace and Facebook. In *Digital Race Anthology* (Eds. Lisa Nakamura and Peter Chow-White). Routledge.
21. boyd, d., Lee, H.Y, Ramage, D., and Donath, J. (2002) Developing legible visualizations for online social spaces. In *Proceedings of the Hawaii International Conference on System Sciences*, Jan. 7–10 2002, Big Island, Hawaii.
22. boyd, d., Golder, S., and Lotan, G. (2010) .Tweet Tweet Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of HICSS-42, Persistent Conversation Track*. Kauai, HI: IEEE Computer Society.
23. Brilliant, R. (1991). *Portraiture*. Cambridge, MA: Harvard University Press.
24. Brown, P. and Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK.
25. Brown, P., Della Pietra, S., Della Pietra, V., Lai, J., Mercer, R. (1992). An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics* 18 (1).
26. Bruckman, A. (1993) Gender swapping on the Internet. *Proceedings of INET'93*.
27. Budiu, R., Pirolli, P.L., Fleetwood, M., Heiser, J. (2006). Navigation in degree of interest trees. 8th International Working Conference on Advanced Visual Interfaces (AVI 2006). May 23-26, Venice, Italy. pp. 457-462. ACM.
28. Burleson, B. R. (1992) On the analysis and criticism of arguments: some theoretical and methodological considerations. In W.L. Benoit, D. Hample and P.J. Penoit (eds), *Readings in Argumentation*. Berlin: Foris.
29. Burke, M., Marlow, C., and Lento, T. (2009) Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems* (Boston, MA, USA, April 04 - 09, 2009). CHI '09. ACM, New York, NY, 945-954. DOI= <http://doi.acm.org/10.1145/1518701.1518847>
30. Byron L. and Wattenberg, B. (2008). *Stacked Graphs - Geometry & Aesthetics*. InfoVis 2008.
31. Carnegie, D. (1936). *How to Win Friends and Influence People*. Simon and Schuster.
32. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*
33. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems*.
34. Clark, H. H. and Brennan, S. E. (1991) Grounding in Communication. In Baecker, R. M. (eds.). *Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration*. San Francisco, CA, USA, Morgan Kaufmann, pp. 222-233.
35. Clark, H.H. (1996) *Using Language*. Cambridge University Press, Cambridge, UK.
36. Cockburn, A., and McKenzie, B. (2000). An Evaluation of Cone Trees. In *People and Computers XIV: British Computer Society Conference on Human Computer Interaction 2000*, pp. 425-436. Springer-Verlag.
37. comScore. (2011). comScore Media Metrix Ranks Top 50 U.S. Web Properties for January 2011. http://www.comscore.com/Press_Events/Press_Releases/2011/2/comScore_Media_Metrix_Ranks_Top_50_U.S._Web_Properties_for_January_2011.
38. Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, Vol. 34, No. 2 (Summer, 2000), pp. 213-238

39. Dave, K., Wattenberg, M., Muller, M. (2004). Flash Forums and ForumReader: Navigating a New Kind of Large-scale Online Discussion. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work 2004 (CSCW'04)*, November 6-10, 2004, Chicago, IL.
40. Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*. Volume 51, Issue 1 (January 2008), pp. 107-113.
41. DeCamp, P., Frid-Jimenez, A., Guinness, J. and Roy, D. (2005). Gist Icons: Seeing Meaning in Large Bodies of Literature. *IEEE Information Visualization Conference*.
42. Dingledine, R., Mathewson, N., and Syverson, P. (2004) Tor: the second-generation onion router. In *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13* (San Diego, CA, August 09 - 13, 2004). *USENIX Security Symposium*. USENIX Association, Berkeley, CA, p21-21.
43. DiMaggio, P., Hargittai, E., Newman, R. W., and Robinson, J. P. (2001). Social Implications of the Internet. *Annual Review of Sociology*, vol 27, pp. 307-36, 2001.
44. DiMicco, J.M., Pandolfo, A., and Bender W. (2004) Influencing group participation with a shared display. In *Proceedings of the 2004 ACM conference on Computer Supported Cooperative Work: CSCW'04*, Chicago, IL, pp. 614-623.
45. Donath, J. (1998) Identity and Deception in the Virtual Community. In *Communities in Cyberspace*. (M. Smith and P. Kollock, eds.) London: Routledge.
46. Donath, J. (2001). Mediated Faces. *Proceedings of the 4th International Conference on Cognitive Technology: Instruments of Mind*. Beynon, M., Nehaniv, C.L., Dautenhahn, K. (Eds.).
47. Donath, J. (2006) The Rhythms of Salience: A Conversation Map. In Abrams, J. and Hall, P. (eds.), *Else/Where: Mapping — New Cartographies of Networks and Territories*. University of Minnesota Design Institute Press.
48. Donath, J. (2007). Virtually trustworthy. *Science*, 317(5834):53-4.
49. Donath, J., Karahalios, K., and Viégas, F. (1999) Visualizing Conversation. In *Proceedings of HICSS-32*, reprinted in the *Journal of Computer Mediated Communication*, vol 4, issue 4.
50. Donath, J., Karahalios, K., and Viegas, F. (2000) Visiphone. *Proceedings of ICAD2000*.
51. Donath, J., Lee Y.H., boyd d, Goler, J. (2001). Loom2: Intuitively Visualizing Usenet. *CSCW 2001 Workshop*.
52. Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification* (2nd ed). Wiley-Interscience.
53. Eagle, N. (2004). Can Serendipity Be Planned? *MIT Sloan Management Review*, Vol. 46, No. 1, pp 10-14.
54. Eick S, Steffen JL, Sumner EE. (1992). Seesoft - A tool for visualizing line oriented software statistics. In *Transactions on Software Engineering*, volume 18, pages 957-968, November, 1992.
55. Ellson, J., Gansner, E., Koutsofios, E., North, S., and Woodhull, G. (2003). Graphviz and dynagraph – static and dynamic graph drawing tools. In *Graph Drawing Software*, M. Junger and P. Mutzel, Eds. Springer- Verlag, Berlin, 127–148.
56. Erickson, T., Halverson, C., Kellogg, W.A., and Laff, M. (2002) Social translucence: designing social infrastructures that make collective activity visible. *Communications of the ACM*, vol. 45, no. 4, pp. 40-44.
57. Erkan, G. and Radev, D.R.. (2004). Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, vol. 22.

58. Ekman, P., & Keltner, D. (1996). Universal facial expressions of emotion: An old controversy and new findings. In U. Segerstrale and P. Molnar (Eds.), *Nonverbal Communication: Where Nature Meets Culture*. (pp. 27-46) NJ: Lawrence Erlbaum Associates.
59. Ennals, R., Hirsch, T., Rattenbury, T., Agost, J.M. (2009) *Dispute Finder*. http://confront.intel-research.net/Dispute_Finder.html
60. Erickson, T., Huang, W., Danis, C. and Kellogg, W. A. (2004). Social Proxy for Distributed Tasks: Design and Evaluation of a Working Prototype. In *Proceedings of CHI'04 Human Factors in Computing Systems*, ACM/SIGCHI, NY, pp. 559-566.
61. Farzan, R., DiMicco, J.M., Millen, D.R., Brownholtz, B., Geyer, W., and Dugan, C. (2008) When the experiment is over: Deploying an incentive system to all the users. *Symposium on Persuasive Technology*, In conjunction with the AISB 2008 Convention, Aberdeen, Scotland, April 2008.
62. Froehlich J, Dourish P. (2004). Unifying Artifacts and Activities in a Visual Tool for Distributed Software Development Teams. *Proceedings of the International Conference on Software Engineering ICSE 2004* (Edinburgh, UK), 387-396.
63. Froehlich, J. (2004). Unifying Artifacts and Activities in a Visual Tool for Distributed Software Development Teams. Master's Thesis. Donald Bren School of Information and Computer Sciences, University of California, Irvine. June 2004.
64. Fry, B. (2003). *Revisionist: Visualizing the evolution of software projects*. <http://acg.media.mit.edu/people/fry/revisionist/>
65. Fry, B. (2004). *Computational Information Design*. PhD Thesis, Massachusetts Institute of Technology, 2004.
66. Galloway, A., Tribe, M., Wattenberg, M. (1999) *StarryNight*. <http://rhizome.org/starrynight/>
67. Gansner, E., Koren, Y. (2007). Improved Circular Layouts. *Lecture Notes in Computer Science, Graph Drawing*, vol 4372/2007, pp. 386-398, Springer Berlin / Heidelberg.
68. Gansner, E., Koren, Y., North, S. (2004). Topological Fisheye Views for Visualizing Large Graphs. In *Proceedings of IEEE Visualization Conference*, IEEE (2004), pp. 175- 182.
69. Garton, L., Haythornthwaite, C., and Wellman, B. (1997) Studying online social networks. *Journal of Computer-Mediated Communication*, Volume 3 Issue 1 June 1997.
70. Gilbert, E, Karahalios, K. (2009) Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, United States). CHI '09. ACM, New York, NY, pp. 211-220.
71. Gleave, E., Welser, H. T., Lento, T. M., and Smith, M. A. (2009). A Conceptual and Operational Definition of 'Social Role' in Online Community. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, IEEE, 2009.
72. Grice, H. P. (1957) Meaning. *Philosophical Review*, vol. 64, pp. 377-388.
73. Grice, H. P. (1969) Utterer's meaning and intentions. *Philosophical Review*, vol. 78, pp. 147-177.
74. Goffman, E. (1959) *The Presentation of Self in Everyday Life*. Anchor Books.
75. Goldberg, K., de Kosnik, G., Ryokai, K., Laslocky, M., Nathanson, T., Bitton, E., Blas, Z., Goodman, L., Faridani, S., Wong, D., and Sydell, A. (2009). *Opinion Space*. <http://opinion.berkeley.edu>
76. Golder, S. A. (2003) *A Typology of Social Roles in Usenet*. Unpublished Senior Honors, Harvard University, Cambridge, MA.
77. Golder, S. (2005) *Webbed Footnotes: Collaborative Annotation on the Web*. Masters Thesis, MIT Media Lab.

78. Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. In Proceedings of the National Academy of Sciences, 2004.
79. Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J. (2005). Integrating topics and syntax. *Neural Information Processing Systems* 17, 2005.
80. Halevy, A., Norvig, P., and Pereira, F. (2009) The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*.
81. Hancock, J., Thom-Santelli, J., and Ritchie, T. (2004) Deception and design: The impact of communication technology on lying behavior. Proceedings of the SIGCHI conference on Human factors in computing systems, Vienna, Austria, April 24-29, 2004.
82. Hancock, J., Toma, C., and Ellison, N. (2007). The Truth about Lying in Online Dating Profiles. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, San Jose, CA, pp. 449-452.
83. Handy, C. (1995). Trust and the virtual organization. *Harvard Business Review*, 7(3):40-50.
84. Hannemann, J. and Kiczales, G. (2001). Overcoming the Prevalent Decomposition of Legacy Code, Workshop on Advanced Separation of Concerns at the International Conference on Software Engineering (ICSE) 2001, Toronto, Canada.
85. Hansell, S. (2008). Zuckerberg's Law of Information Sharing. *The New York Time Bits Blog*, November 6, 2008.
86. Hassan-Montero, Y. and Herrero-Solana, V. (2006) Improving tag-clouds as visual information retrieval interfaces. In Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006. Mérida, Spain, October 25-28, 2006.
87. Havre, S., Hetzler, E., Whitney, P., and Nowell, L. (2002). ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (January 2002), 9-20. DOI=10.1109/2945.981848 <http://dx.doi.org/10.1109/2945.981848>
88. Hearst, M. A., & Rosner, D. (2008). Tag Clouds: Data Analysis Tool or Social Signaller In Hawaii International Conference on System Sciences, Proceedings of the 41st Annual (p. 160). Presented at the Hawaii International Conference on System Sciences, Proceedings of the 41st Annual. doi: 10.1109/HICSS.2008.422
89. Herring, S. (1999) Interactional Coherence. *Journal of Computer-Mediated Communication*, vol. 4, no. 4.
90. Hiltz, S. R., and Turoff, M. (1978) *The network nation: Human communication via computer*. Reading, Massachusetts: Addison Wesley.
91. Himma, K. E. (2007) A Preliminary Step in Understanding the Nature of a Harmful Information-Related Condition: An Analysis of the Concept of Information Overload. *Ethics and Information Technology*, Vol. 9, No. 4, December 2007. Available at SSRN: <http://ssrn.com/abstract=954719>
92. Holland, J., and Stornetta, S. (1992) Beyond being there. In proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Monterey, California), ACM Press, New York, pp. 119-125.
93. Hudson, R. (1996) *Sociolinguistics* (2nd ed), Cambridge University Press.
94. Ito, M., Okabe, D., Matsuda, M. (2005) *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life*. The MIT Press, Cambridge, MA.

95. Johnson, B., and Shneiderman, B. (1991). Treemaps: A SpaceFilling Approach to the Visualization of Hierarchical Information Structures. *Proc. 2nd International IEEE Visualization Conference*, IEEE (1991), 284-291.
96. Jones, J., Harrold, .M, Stasko J. (2001). Visualization for Fault Localization. *Proceedings of the Workshop on Software Visualization, 23rd International Conference on Software Engineering*, Toronto, Ontario, Canada.
97. Jones, S., and Fox, S. (2009) *Generations Online in 2009*. Pew Internet & American Life Project, January 2009.
98. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P. and Van Ess-Dykema, C. (1997a) Switchboard discourse language modeling project report. Technical report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD.
99. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P. and Van Ess-Dykema, C. (1997b) Automatic detection of discourse structure for speech recognition and understanding. *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, pp. 88-95.
100. Khan, F., Fisher, T., Shuler, L., & Wu, T. (2002). Mining chat-room conversations for social and semantic interactions.
101. Kiesler, S., Siegel, J., Mcguire, T. W. (1984) Social psychological aspects of computer-mediated communication. *American Psychologist*, vol. 39 (10), October 1984.
102. Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*. 73: 31-68
103. Kittur, A., Suh, B., Pendleton, B.A., and Chi, E.H. (2007) He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, San Jose, CA, pp. 453-462.
104. Kim, T., and Pentland, A. (2009). Understanding Effects of Feedback on Group Collaboration. *Human Behavior Modeling, AAAI Spring Symposium*. Palo Alto, CA. March 2009.
105. Klingberg, T. (2008) *The overflowing brain: information overload and the limits of working memory*. Oxford University Press.
106. Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), 2002.
107. Krauss, R.M., and Fussell, S.R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, vol. 9, pp. 2-24.
108. Kraut, R.E., Gergle, D. and Fussell, S.R. (2002). The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-Presence. *CSCW'02*, pp. 31-40.
109. Krishnamurthy S. (2002). *Cave or Community? An Empirical Examination of 100 Mature Open Source Projects*. University of Washington. May 2002.
110. Kruskal, J.B, and Landwehr, J.M. (1983). Icicle Plots: Better Displays for Hierarchical Clustering. *The American Statistician*, vol 37, no 2. pp. 162-168. American Statistical Association. 1983.
111. Kubat, R., DeCamp, P., Roy, P., and Roy, D. (2007). TotalRecall: Visualization and Semi-Automatic Annotation of Very Large Audio-Visual Corpora. *Ninth International Conference on Multimodal Interfaces (ICMI 2007)*.
112. Kunda, Z. (1999) *Social cognition: making sense of people*. MIT Press, Cambridge, MA.

113. Lakhani K, Wolf R. (2005). *Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects*. Perspectives on Free and Open Source Software. MIT Press, Cambridge, MA.
114. Lakoff, G. (1987). *Woman, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press.
115. Lam, F. and Donath, J. (2005) Seascape and volcano: visualizing online discussions using timeless motion. In *Proceedings of CHI '05 extended abstracts, Conference on Human factors in Computing Systems*, Portland, OR, pp. 1585-1588.
116. Lamping, L., and Rao, R. (1996). The Hyperbolic Browser: A Focus+Context Technique for Visualizing Large Hierarchies. *Journal of Visual Languages & Computing* Volume 7, Issue 1, pp. 33-55.
117. Lamont, M. (1992). *Money, Morals and Manners: The Culture of the French and the American Upper-Middle Class*. University of Chicago Press
118. Lampe, Cliff and Paul Resnick. (2004) Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems*, Vienna Austria.
119. Lanier, J. (2010) *You Are Not a Gadget*. Alfred A. Knopf Press, New York.
120. Lea, M., and Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing*. (2,3,4) 321-341.
121. Lea, M., and Spears, R. (1995). Love at first byte? building personal relationships Over computer networks. In J. T. Wood and S. Duck (Eds.), *Under-studied Relationships: Off the Beaten Track*. (pp. 197-233) Thousand Oaks, CA: Sage Publications.
122. Leskovec J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. ACM, New York, NY, USA, 497-506.
123. Lessig, L. (2006). *Code: And Other Laws of Cyberspace, Version 2.0*. Basic Books.
124. Liakopoulos, M. (2000). *Argumentation Analysis. Qualitative Researching with Text, Image and Sound*. Bauer, M. and Gaskell, G. (eds.). Sage, London.
125. Li, W., and McCallum, A. (2006) Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of ICML 2006*.
126. Lin, C., and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada.
127. Liu, H. (2007) Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1), Blackwell Publishing.
128. Liu, S., Zhou, M.X., Pan, S., Qian, W., Cai, W., and Lian, X. (2009). Interactive, Topic-based Visual Text Summarization and Analysis. *CIKM '09*, November 2-6, 2009, Hong Kong, China. ACM.
129. Lohr, S. (2007) Is Information Overload a \$650 Billion Drag on the Economy? *New York Times BITS Blog*, December 20, 2007. <http://bits.blogs.nytimes.com/2007/12/20/is-information-overload-a-650-billion-drag-on-the-economy>
130. Lü, H. and Fogarty, J. (2008). Cascaded Treemaps: Examining the Visibility and Stability of Structure in Treemaps. *Graphics Interface Conference 2008*. May 28-30, Windsor, Ontario, Canada. ACM.

131. Lux, M., Granitzer, M., & Kern, R. (2007). Aspects of Broad Folksonomies. Database and Expert Systems Applications, 2007. DEXA '07. 18th International Workshop on, 283–287. doi:10.1109/DEXA.2007.80
132. Mackinlay, J. D. (1986). Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics, 5. pp. 110-141.
133. Maletic, J., Marcus, A, and Feng, L.(2003). Source Viewer 3D (sv3D)-A Framework for Software Visualization. ICSE 2003, Portland, IEEE CS Press, May 2003.
134. Malone, T. W., Grant, K. R., Lai, K., Rao, R., and Rosenblitt, D. (1987) Semistructured messages are surprisingly useful for computer-supported coordination. ACM Trans. Inf. Syst. vol. 5, no. 2 (Apr. 1987), 115-131. DOI= <http://doi.acm.org/10.1145/27636.27637>
135. Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
136. Matsuda, K., Miyake, T., and Kawai, H. (2002) Culture formation and its issues in personal agent-oriented virtual society: "PAW^2". In Proceedings of the 4th international Conference on Collaborative Virtual Environments (Bonn, Germany, September 30 - October 02, 2002). CVE '02. ACM, New York, NY, 17-24. DOI= <http://doi.acm.org/10.1145/571878.571882>
137. Marwick, A. and boyd, D. (2011). I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. New Media and Society, 13, pp. 96-113.
138. McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>.
139. McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of Artificial Intelligence Research (JAIR)
140. McCarthy, C. Facebook F8: One graph to rule them all. cnet News, April 21 2010. http://news.cnet.com/8301-13577_3-20003053-36.html
141. McCarthy, J. C., Miles, V. C., and Monk, A. F. (1991) An experimental study of common ground in text-based communication. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology (New Orleans, Louisiana, United States, April 27 - May 02, 1991). ACM, New York, NY, 209-215. DOI= <http://doi.acm.org/10.1145/108844.108890>
142. McLuhan, M., and Fiore, Q. (1967). The Medium is the Massage: An Inventory of Effects. Bantam Books/Random House.
143. Messmer, E. (2011). Symantec: Cybercriminals use social networking sites as preferred form of attack. TechWorld, April 11th, 2011, <http://news.techworld.com/security/3272456/symantec-cybercriminals-use-social-networking-sites-as-preferred-form-of-attack/>
144. Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.
145. Mihalcea, R. and Strapparava, C. (2006). Learning to Laugh (Automatically): Computational Models for Humor Recognition. Journal of Computational Intelligence.
146. Mills, S. (1869). On Liberty. London: Longman, Roberts & Green, 1869. New York: bartleby.com, 1999.
147. Mimno, D., and McCallum, A. (2008) Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. Conference on Uncertainty in Artificial Intelligence (UAI '08).

148. Mishne, G., and Glance, N. (2006) Leave a Reply: An Analysis of Weblog Comments. Proceedings of WWW2006, May 22-26, 2006, Edinburgh, UK.
149. Moreno, J. (1934). Who Shall Survive? Beacon House Inc. Beacon, N.Y.
150. Mukherjee, A., and Liu, B. (2010). Improving gender classification of blog authors. In EMNLP '10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
151. Nardi, B. A., Whittaker, S., Isaacs, E., Creech, M., Johnson, J., and Hainsworth, J. (2002). Integrating communication and information through ContactMap. *Communications of the ACM*, vol. 45, no. 4, pp. 89-95. DOI= <http://doi.acm.org/10.1145/505248.505251>
152. Newman, M., Pennebaker, J., Berry, D., and Richards, J. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, (23), pp. 665.
153. Norman, D. (1998) *The Design of Everyday Things*. MIT Press, Cambridge, MA.
154. Offenhuber, D., and Donath, J. (2008) Comment Flow: Visualizing Communication Along Network Path. In Sommerer, C., Mignonneau, L. & King, D. (eds.), *Interface Cultures: Artistic Aspects of Interaction* 1st ed., Transcript.
155. On, J. (2004) They Rule. <http://www.theyrule.net>
156. Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79–86.
157. Pariser, E. (2011). *The filter bubble: what the Internet is hiding from you*. New York: Penguin Press.
158. Park, R. E. (1950). *Race and Culture*. The Free Press, Glencoe, Illinois.
159. Pekacki, L., Cyganiak, R., Jentsch, A., Schulze, M., Pingel, S., and van Lessen, T. (2009). StatCVS. <http://statcvs.sourceforge.net/>
160. Peña, J. and Hancock, J. T. (2006). An analysis of instrumental and socio-emotional content in online multi-player videogames. *Communication Research*. 33:92-109.
161. Perry, E. (2004). *Anthropomorphic Visualization: Depicting Participants in Online Spaces Using the Human Form*. M.S. Thesis, Media Arts and Sciences, MIT Media Lab. Cambridge, MA.
162. Pirolli, P., Card, S.K., and Van Der Wege, M.M. (2000). The effect of information scent on searching information: visualizations of large tree structures. In *Proceedings of the Working Conference on Advanced Visual Interfaces (Palermo, Italy)*. AVI '00. ACM, New York, NY, 161-172.
163. Plaisant, C., Grosjean, J., and Bederson, B. (2002). SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation. In *Proceedings of the IEEE Symposium on information Visualization (Infovis'02)* IEEE Computer Society, Washington, DC.
164. Preece, J. (2001) Sociability and usability: Twenty years of chatting online. *Behavior and Information Technology Journal*, vol. 20, no. 5, pp. 347-356.
165. Preece, J. and Diane Maloney-Krichmar (2003) Online Communities: Focusing on sociability and usability. In J. Jacko and A. Sears, A. (Eds.) *Handbook of Human-Computer Interaction*, Lawrence Erlbaum Associates Inc. Publishers. Mahwah, NJ. pp. 596-620.
166. Radev, D., Jin, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*.

167. Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Center for the Study of Language and Information, Univ. of Chicago Press.
168. Reid, E. (1994) *Cultural formations in text-based virtual realities*. Thesis, Dept. of English, University of Melbourne.
169. Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.
170. Robertson, G., Mackinlay, J.D., and Card, S. (1991). Cone Trees: Animated 3D Visualizations of Hierarchical Information. In *Proceedings of the ACM CHI '91 Human Factors in Computing Systems Conference*, pages 189-194, April 28 - June 5, 1991, New Orleans, Louisiana. ACM.
171. Ramage, D., Dumais, S., and Liebling, D. (2010). Characterizing Microblogs with Topic Models. In *Proceedings of ICWSM 2010*.
172. Rogowitz, B.E. & Treinish, L.A. (1996). How NOT to Lie with Visualization. *Computers in Physics*, 10(6):268–273.
173. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence*.
174. Ruch, W. (1998). *The sense of humor: explorations of a personality characteristic*. Mouton de Gruyter.
175. Sack, W. (2001). Conversation Map: An Interface for Very Large-Scale Conversations. *Journal of Management Information Systems*, Vol. 17 No. 3, Winter 2001 pp. 73 - 92.
176. Sack, W. (2007). *Aesthetics of Information Visualization*. Paul, C., Vesna, V., and Lovejoy, M. (eds.) Context Providers, University of Minnesota Press.
177. Simmel, G. (1910). How is society possible? *American Journal of Sociology*, vol 16, pp. 372-391, 1910.
178. Simmel, G. (1950). *The Stranger*. From Kurt Wolff (Trans.) *The Sociology of Georg Simmel*. Free Press, New York, pp. 402 - 408, 1950.
179. Singh, L., and Zhan, J. (2007). Measuring topological anonymity in social networks. *Intl. Conf. on Granular Computing*.
180. Shahaf, D. and Amir, E. (2007). Towards a theory of AI completeness. *Commonsense 2007*, 8th International Symposium on Logical Formalizations of Commonsense Reasoning.
181. Smith, A., Schlozman, K. L., Verba, S., and Brady, H. (2009) *The Internet and Civic Engagement*. Pew Internet & American Life Project, Pew Research Center, Washington, D.C. September 2009.
182. Smith, M. (1998) *Invisible crowds in cyberspace: Mapping the social structure of the Usenet*. In *Communities in Cyberspace*. (M. Smith and P. Kollock, eds.) London: Routledge.
183. Smith, M. A. and Fiore, A. T. 2001. Visualization components for persistent conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, United States). CHI '01. ACM, New York, NY, 136-143. DOI= <http://doi.acm.org/10.1145/365024.365073>
184. Smith, M. and Kollock, P. (1996) *Managing the Virtual Commons: Cooperation and Conflict in Computer Communities*. In S. Herring (eds.), *Computer-Mediated Communication*, John Benjamins, Amsterdam.
185. Smith, M., Cadiz, J. J., and Burkhalter, B. (2000) *Conversation trees and threaded chats*. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative*

- Work (Philadelphia, Pennsylvania, United States). CSCW '00. ACM, New York, NY, 97-105. DOI= <http://doi.acm.org/10.1145/358916.358980>
186. Soroker, D., Zinman, A., and Narayanaswami, C. (2008) Organizational Maps and Mashups. IBM Technical Report RC24551, Watson Research Center.
 187. Sproull L., and Kiesler, S. (1986) Reducing social context cues: Electronic mail in organizational communications. *Management science*, vol 32, no 11, pp. 1492-1512, November 1986.
 188. Sproull, L., Kiesler, S., and Zubrow, D. (1984). Encountering an alien culture. *Journal of Social Issues*, v40 n3 p31-48 1984
 189. Stasko, J., and Zhang, E. (2000). Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the IEEE Symposium on Information Visualization 2000*, pp.57- 65.
 190. Steyvers, M. and Griffiths, T. (2005). Matlab Topic Modeling Toolbox. http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
 191. Sussman, S. and Sproull, L. (1999) Straight talk: Delivering bad news through electronic communication. *Information Systems Research* vol. 10 (2) pp. 150-166, 1999.
 192. Thornton, S. (1996). *Club Cultures: Music, Media, and Subcultural Capital*. University Press of New England, Hanover, N.H.
 193. Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society*. 4th edition, Penguin Books, London, UK.
 194. Toulmin, S. (1969) *The Uses of Argument*, Cambridge, England: Cambridge University Press.
 195. Turner, T.C., Smith M., Fisher, D., Welser H.T. (2005) Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer Mediated Communication*, vol. 10, no. 4.
 196. Twitter. (2011). #numbers. <http://blog.twitter.com/2011/03/numbers.html>
 197. Urbandictionary. (2011). AzN. <http://www.urbandictionary.com/define.php?term=azn>, downloaded June 6th, 2011.
 198. Viégas, F. (2005) *Revealing individual and collective pasts: Visualizations of online social archives*. PhD Thesis, MIT Media Lab.
 199. Viégas, F., boyd, d., Nguyen, D.H., Potter, J., and Donath, J. (2004) Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments. In *Proceedings of HICSS-37*, Hawaii, HI. January 5-8, 2004.
 200. Viégas, F., Smith, M. (2004) Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces. In *Proceedings of HICSS-37*, Hawaii, HI. January 5-8, 2004.
 201. Viégas, F., Perry, E., Howe, E., and Donath, J. (2004). Artifacts of the Presence Era: Using Information Visualization to Create an Evocative Souvenir. *InfoVis2004*. 10-12 October, 2004. Austin, TX.
 202. Viégas, F., Wattenberg, M., van Ham, F., Kriss, J. McKeon, M. (2007). ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, Volume: 13 (6), pp. 1121-1128.
 203. Viégas F, Wattenberg M, Kushal D. (2004). Studying Cooperation and Conflict between Authors with history flow Visualizations. *CHI 2004*. Vienna, Austria. 24-29 April, 2004.
 204. Wattenberg, M. and Millen, D.R. (2003) Conversation thumbnails for large-scale discussions. In *Proceedings of CHI '03: Conference on Human Factors in Computing Systems*, Ft. Lauderdale, FL. pp. 742-743

205. Watts, R.J. (2003). *Politeness*. Cambridge University Press, Cambridge UK.
206. Webster, J., and Trevino, L. K. (1995) Rational and social theories as complementary explanations of communication media choices: Two policy-capturing studies. *Academy of Management Journal*. vol. 38 (6) pp. 1544-1572, 1995
207. Wellens, A.R. (1986). Use of a psychological distancing model to assess differences in telecommunication media. In L. Parker & C. Olgren (Eds.), *Teleconferencing and electronic media*. Vol. V. Madison, Wisconsin: Center for Interactive Programs, University of Wisconsin.
208. Welser, H.T., Gleave, E., Fisher, D., and Smith, M. (2007) Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, vol. 8.
209. Welser, H. T., Kossinets, G. , Smith, M. A. and Cosley, D. (2008) Finding social roles in Wikipedia. Paper presented at the annual meeting of the American Sociological Association Annual Meeting, Sheraton Boston and the Boston Marriott Copley Place, Boston, MA.
210. Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19, 52-90.
211. Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, p.735-740.
212. West, M. (1953). *A General Service List of English Words*. Longman, London.
213. Williams, C. (2011). How Egypt shut down the internet. *The Telegraph*. Jan 28, 2011. <http://www.telegraph.co.uk/news/worldnews/africaandindianocean/egypt/8288163/How-Egypt-shut-down-the-internet.html>
214. Wise, J. A. , Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., and Schur, A. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. *Proceedings of Information Visualization Symposium '95*, Gershon, N. and Eick, S.G. (eds.), Computer Society Press, Los Alamitos, CA, pp. 51–58.
215. Wiseman, R. (2002). *Queen Bees and Wannabes: Helping Your Daughter Survive Cliques, Gossip, Boyfriends, and Other Realities of Adolescence*. Crown.
216. Whittaker, S. (2002) Theories and Methods in Mediated Communication. In Graesser, A., Gernsbacher, M., and Goldman, S. (Ed.) *The Handbook of Discourse Processes*, 243-286, Erlbaum, NJ. Whittaker, S., Bellotti, V., and Gwizdka, J. (2006) Email in personal information management. *Commun. ACM* 49, 1 (Jan. 2006), pp. 68-73. DOI= <http://doi.acm.org/10.1145/1107458.1107494>
217. Whittaker, S., Terveen, L., Hill, W., and Cherny, L. (1998) The dynamics of mass interaction. *Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work* (Seattle, WA), pp. 257-264.
218. Whyte, W. (1980) *The Social Life of Small Urban Spaces*. Washington, D.C.: The Conservation Foundation.
219. Winograd, T. (1987) A Language/Action Perspective on the Design of Cooperative Work. In *Human-Computer Interaction*, Vol 3 No. 1, pp. 3-30.
220. Xiong, R. and Donath, J. (1999) PeopleGarden: creating data portraits for users. *Proceedings of UIST '99, the 12th annual ACM symposium on User interface software and technology*, Asheville, N.C. pp. 37-44.
221. Yee, N. (2007). Motivations of Play in Online Games. *Journal of Cyber Psychology and Behavior*, 9, 772-775.

- 222. Zebrowitz, L. (1996). Physical appearance as a basis of stereotyping. In C. N. Macrae, C. Stangor, and M. Hewstone (Eds.), *Stereotypes and Stereotyping*. (pp. 79-120) New York: The Guilford Press.
- 223. Zebrowitz, L. (1997) *Reading Faces*. Westview Press, Boulder, CO.
- 224. Zinman, A. (2004) *Open Sources*. <http://smg.media.mit.edu/projects/OpenSources>
- 225. Zinman, A. (2006) *RadioActive: Enabling Large-Scale Asynchronous Audio Discussions on Mobile Devices*. MS Thesis, Media Arts and Sciences, Massachusetts Institute of Technology.
- 226. Zinman, A., and Donath, J. (2007) *Is Britney Spears Spam?* In *Proceedings of Fourth Conference on Email and Anti-Spam*, Mountain View, California, August 2-3, 2007.
- 227. Zinman A. and Donath, J. (2009) *Signs: Increasing Expression and Clarity in Instant Messaging*. Proc. HICSS.

APPENDIX A: SURVEY RESULTS

Community participation

- Would you say your commenting behavior is not significantly different across all PRODUCT REVIEW sites?
28 (93.33%): Yes, I only want to answer questions about PRODUCT REVIEW sites as a whole
1 (3.33%): No, I want to answer questions about each PRODUCT REVIEW site I list separately
- Would you say your commenting behavior is not significantly different across all NEWS sites?
18 (85.71%): Yes, I only want to answer questions about NEWS sites as a whole
2 (9.52%): No, I want to answer questions about each NEWS site I list separately
- Would you say your commenting behavior is not significantly different across all MAJOR BLOGS?
12 (100.00%): Yes, I only want to answer questions about MAJOR BLOGS as a whole
0 (0.00%): No, I want to answer questions about each MAJOR BLOG I list separately
- Would you say your commenting behavior is not significantly different across all SOCIAL MEDIA?
33 (91.67%): Yes, I only want to answer questions about SOCIAL MEDIA as a whole
1 (2.78%): No, I want to answer questions about each SOCIAL MEDIA site I list separately
- Would you say your commenting behavior is the same across all AGGREGATOR sites?
12 (100.00%): Yes, I only want to answer questions about AGGREGATOR sites as a whole
0 (0.00%): No, I want to answer questions about each AGGREGATOR site I list separately
- Do you ever read and/or write comments in the following categories?
 - 23 -- Aggregators, i.e. Digg/Reddit/Technorati
 - 1 (7.69%): redditandslashdot
 - 4 (30.77%): reddit
 - 5 (38.46%): digg
 - 3 (23.08%): hackernews
 - 10 -- Other
 - 1 (20.00%): craigslistforums
 - 1 (20.00%): make
 - 1 (20.00%): googlereader
 - 1 (20.00%): crossfit.com
 - 1 (20.00%): craigslistwriter'sforum
 - 60 -- Product reviews, i.e. Amazon/Best Buy/Yelp
 - 1 (1.79%): foodtv
 - 1 (1.79%): vitacost
 - 1 (1.79%): costco
 - 3 (5.36%): tripadvisor
 - 12 (21.43%): yelp
 - 1 (1.79%): amazaon
 - 1 (1.79%): bing

- 1 (1.79%): thorlabs
- 1 (1.79%): citysearch
- 24 (42.86%): amazon
- 1 (1.79%): newegg
- 1 (1.79%): amazon&yelp(thoughonlywhenilivedinsf)
- 1 (1.79%): edmundoptics
- 1 (1.79%): googlemaps
- 1 (1.79%): epicurious
- 1 (1.79%): tigerdirect
- 1 (1.79%): blockbuster
- 1 (1.79%): backcountry
- 1 (1.79%): bizrate
- 1 (1.79%): amaz
- 29 -- Media sites, i.e. YouTube/SoundCloud/t61
 - 1 (5.56%): reddit
 - 1 (5.56%): ithinkyoutubeisaboutit.
 - 3 (16.67%): vimeo
 - 11 (61.11%): youtube
 - 1 (5.56%): imdb
 - 1 (5.56%): flickr
- 23 -- Major blogs, i.e. Huffington Post/BoingBoing/Politico/Slashdot
 - 3 (15.00%): boingboing
 - 2 (10.00%): politico
 - 1 (5.00%): mostlikelyhuffingtonpost.
 - 1 (5.00%): dailycaller
 - 1 (5.00%): huffingtonpostblogger
 - 1 (5.00%): engadget
 - 1 (5.00%): huffpost
 - 3 (15.00%): slashdot
 - 2 (10.00%): gizmodo
 - 1 (5.00%): boygeniusreport
 - 1 (5.00%): peoplesendmethingsonthese
 - 2 (10.00%): techcrunch
 - 1 (5.00%): smashingmagazine
- 38 -- News sites, i.e. NYTimes/WSJ/Indymedia
 - 1 (2.78%): marca.com
 - 1 (2.78%): nytimes.com
 - 3 (8.33%): washingtonpost
 - 7 (19.44%): nytimes
 - 1 (2.78%): guardian.co.uk
 - 1 (2.78%): huffingtonpost
 - 2 (5.56%): sfgate
 - 1 (2.78%): many
 - 1 (2.78%): slate
 - 1 (2.78%): fredericknewspost
 - 1 (2.78%): foxnews
 - 2 (5.56%): nyt
 - 1 (2.78%): yahoo
 - 1 (2.78%): wmur
 - 1 (2.78%): salon

2 (5.56%): newyorktimes
 1 (2.78%): nashuatelegraph
 4 (11.11%): cnn
 1 (2.78%): trueslant
 1 (2.78%): salon.com
 2 (5.56%): unionleader
 19 -- Other blogs
 1 (8.33%): somefriends'personalblogs
 1 (8.33%): failblog
 1 (8.33%): miscellaneous
 1 (8.33%): frederickmarylandonline
 1 (8.33%): engadget
 1 (8.33%): facebook.
 1 (8.33%): friend'sblogs
 1 (8.33%): hyperallergic.com
 1 (8.33%): ffffound
 1 (8.33%): art21blog
 1 (8.33%): cookscorner
 1 (8.33%): friendsblogs...
 67 -- Social media, i.e. Facebook/MySpace comments on updates, etc.
 1 (2.50%): facebook;veryrare
 4 (10.00%): twitter
 2 (5.00%): linkedin
 1 (2.50%): tumblr
 1 (2.50%): foursquare
 31 (77.50%): facebook

■ Would you say your commenting behavior is not significantly different across all OTHER BLOGS?

10 (100.00%): Yes, I only want to answer questions about OTHER BLOGS as a whole
 0 (0.00%): No, I want to answer questions about each OTHER BLOG site I list separately

■ Roughly speaking, how many comments do you think you EVER written?

2 (4.08%): Zero
 4 (8.16%): A few
 4 (8.16%): ~10
 5 (10.20%): 10-25
 8 (16.33%): 25-50
 9 (18.37%): 50-100
 9 (18.37%): 100-500
 4 (8.16%): 500-1000
 4 (8.16%): 1000+

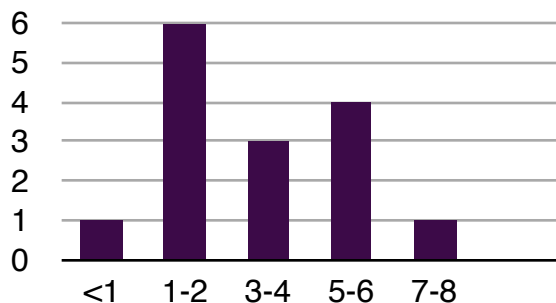
■ Would you say your commenting behavior is not significantly different across all MEDIA sites?

15 (100.00%): Yes, I only want to answer questions about MEDIA sites as a whole
 0 (0.00%): No, I want to answer questions about each MEDIA site I list separately

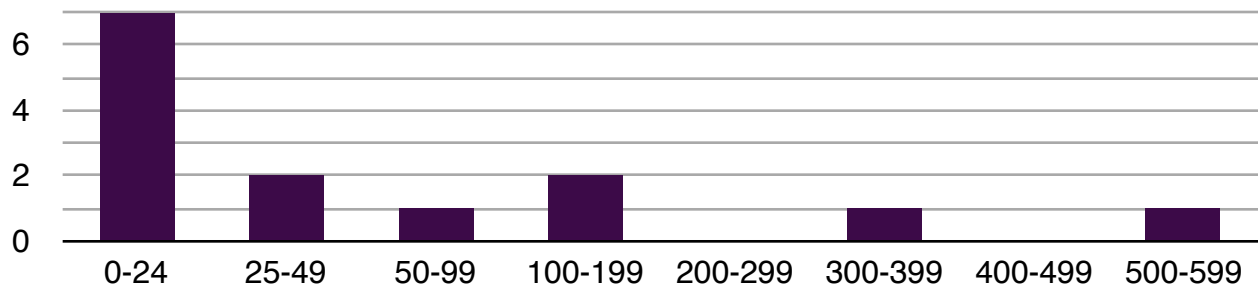
Site Owners

BASICS

- Is the number of writers increasing over time?
4 (26.67%): Yes
11 (73.33%): No
- Are the number of readers increasing over time?
9 (60.00%): Yes
6 (40.00%): No
- What kind of site is it?
0 (0.00%): Aggregators, i.e. Digg/Reddit/Technorati
2 (13.33%): Other
0 (0.00%): Product reviews, i.e. Amazon/Best Buy/Yelp
0 (0.00%): Media sites, i.e. YouTube/SoundCloud/t61
1 (6.67%): Major blogs, i.e. Huffington Post/BoingBoing/Politco/Slashdot
0 (0.00%): News sites, i.e. NYTimes/WSJ/Indymedia
10 (66.67%): Other blogs
2 (13.33%): Social media, i.e. Facebook/MySpace comments on updates, etc.
- How old is your site?



- How many people do you estimate read comments in your site on a daily basis?



- Was your commenting system created by you?
0 (0.00%): Yes

15 (100.00%): No

- What purposes does commenting serve to you, the site owner?

Reader interaction

Interaction with my readers

It allows me to get feedback from my readers and engage with them on new ideas.

discussion

Keeping in touch, entertainment

Engage in discussion/promotion with the community.

Very little as I don't have much traffic

helps to engender a discussion around what i am writing about, helps to build a community

This site is for professional purposes, for communicating ideas to my client. The client's comments tell me whether I am working in directions that they like.

It is an easy way to keep up with friends that I don't necessarily have the time or interest in talking with on the telephone.

Lets people get in touch with me

Replying to other people's comments. Building a fan base

Shows me that people are reading it and have insight into my comments

Mostly for friends to give feedback

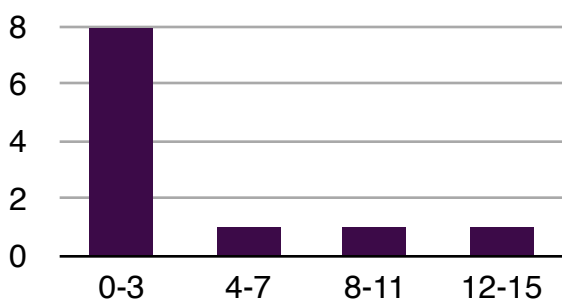
just a default setting, i actually rarely post, so I rarely read comments or maintain the site.

- Has there been a dramatic change in user behavior since adding comments?

0 (0.00%): Yes

2 (100.00%): No

- How many minutes do people spend in using the comments feature on average?



- How important are comments to your site?

2 (14.29%): Essential

3 (21.43%): Very important

4 (28.57%): Important

3 (21.43%): Somewhat important

2 (14.29%): Not important

- What purposes does commenting serve to users?

It allows them to chat with me about recent posts and ideas.

discussion

same as above

Promotion and QA.

gives them an outlet to respond to something I read

Interacting with writers

Keeping in touch, entertainment

The site is a conversation. The comments are their half of the conversation.

It allows them to keep in touch with me. It also allows them to interact with me.

To help them feel that they are part of a dialog

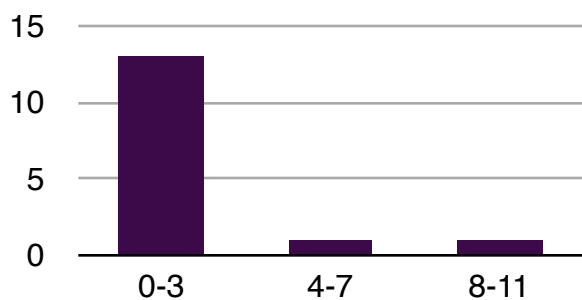
gives them a chance to agree or disagree with the reviews

- Have you always had comments on your site?

13 (86.67%): Yes

2 (13.33%): No

- How many people do you estimate write comments in your site on a daily basis?



- If you found a better commenting platform, would you make the effort to switch?

9 (60.00%): Yes

6 (40.00%): No

- How happy are you with your commenting solution?

2 (13.33%): Very

9 (60.00%): Mostly

4 (26.67%): Somewhat

0 (0.00%): A bit

0 (0.00%): Not at all

NEW FEATURES

- Would exposing relationships address any existing pain points?

Yes

Yes.

many commenters cannot reveal real identities due to official capacity; they are posting as private citizens

Would dynamic surveys address any existing pain points?

Yes

No.

no

microvoting is cool, and seems fun

- How much would you like for the types of comments to be somehow distinguished in the interface? Comments could be grouped along these lines, or at least be color coded.

16 (13.11%): Yes, Please!

41 (33.61%): Sounds good

34 (27.87%): Sure, why not

18 (14.75%): Maybe

13 (10.66%): No thank you

- How much would you like for authors to have a visual overview of their post history next to their comments?

19 (15.32%): Yes, Please!

36 (29.03%): Sounds good

38 (30.65%): Sure, why not

18 (14.52%): Maybe

13 (10.48%): No thank you

- Would comment type discrimination address any existing pain points?

Yes

Hard to visually search through long comment threads

no

No

No.

- Would exposing history address any existing pain points?

If there is a long string of comments it might help address questions

No.

No

Yes

yes

- Would having users structure the comments address any existing pain points?

Yes

No.

yes

No

maybe, maybe not.

maybe, maybe not.

they could more easily sort out the different viewpoints and feeling

- How much would you like for authors to structure the discussion space? E.g. tag comments, draw relationships between comments (repeats, counter-arguments, etc), bump up certain comments and diminish others, etc

24 (19.35%): Yes, Please!
33 (26.61%): Sounds good
31 (25.00%): Sure, why not
21 (16.94%): Maybe
15 (12.10%): No thank you

- Would identity tags address any existing pain points?

No
Possibly
No.

- How much would you like for the relationships between commenters to be exposed? E.g. how often times they've referenced each other, how often they post comments on the same page, if they're friends on Facebook, etc.

25 (20.16%): Yes, Please!
27 (21.77%): Sounds good
25 (20.16%): Sure, why not
25 (20.16%): Maybe
22 (17.74%): No thank you

- How much would you like for posters to "tag their identity" with their posts? For example, one might choose to represent *chinese christian communists* for a given post. Posts would then be grouped by identity, in addition to showing the tags for a given post.

8 (6.50%): Yes, Please!
15 (12.20%): Sounds good
23 (18.70%): Sure, why not
33 (26.83%): Maybe
44 (35.77%): No thank you

- How much would you like for commenters to pose questions to other commenters in the form of a simple survey? For example, in an NYTimes article on the World Cup one might ask other commenters which country they hope will win. Their answer could be instead of a freeform text comment, or integrated with it.

19 (15.45%): Yes, Please!
18 (14.63%): Sounds good
33 (26.83%): Sure, why not
32 (26.02%): Maybe
21 (17.07%): No thank you

An average across all site categories

CONSENSUS

- Do you feel that you try to bring others to a common consensus IN THIS SITE/CATEGORY?

8 (23.53%): Yes
26 (76.47%): No

- Would you like the design of the commenting system to help achieve consensus, or at least map out the arguments IN THIS SITE/CATEGORY?

22 (66.67%): Yes

11 (33.33%): No

- Do you think its desirable (or relevant) to attempt consensus within comments IN THIS SITE/CATEGORY?

3 (2.73%): Its the main point

1 (0.91%): Very desirable and achievable

14 (12.73%): If possible

16 (14.55%): Perhaps, but it's unlikely to happen

14 (12.73%): Can't happen

62 (56.36%): Not relevant

INTERACTION

- What motivates you to leave a comment IN THIS SITE/CATEGORY?

27 -- When I see unbalanced opinions/sides

33 -- For fun/entertainment

32 -- When something is funny

28 -- When I disagree with a post

43 -- When I feel connected to what's being discussed

51 -- When I have something unique to say

9 -- Because I'm awesome and others should listen to me

22 -- I bought the product discussed

14 -- I want to warn others

11 -- Other

- When you post IN THIS SITE/CATEGORY, do you first read others' comments?

32 (35.96%): Always

27 (30.34%): Often

27 (30.34%): Sometimes

3 (3.37%): Rarely

0 (0.00%): Never

0 (0.00%): Other

- How many comments do you read before posting IN THIS SITE/CATEGORY?

12 (13.48%): All

20 (22.47%): Most

11 (12.36%): Half

35 (39.33%): Some

11 (12.36%): Few

- What kinds of comments do you leave IN THIS SITE/CATEGORY?

31 -- Emotional responses to the content (e.g. this made me feel angry)

31 -- Opinions about society/humanity

49 -- Replies to other commentors

35 -- Critical analysis

41 -- Let others know of an experience I've had

16 -- Keep the flow of the discussion going

8 -- I was here

16 -- Jokes
6 -- Egging others on (e.g. trolling)
28 -- I liked/disliked/laughed at this, and that's all I have to say
6 -- Other

VIEWPOINTS

- If you could easily see comments grouped by how people think IN THIS SITE/CATEGORY, would that be a preferred way to read the comments?

44 (56.41%): Yes

34 (43.59%): No

- How often do you prefer to read comments from people who think like you IN THIS SITE/CATEGORY?

4 (3.74%): Always

29 (27.10%): Often

38 (35.51%): Sometimes

5 (4.67%): Rarely

3 (2.80%): Never

23 (21.50%): Not applicable

5 (4.67%): Other

NEW FEATURES

- How much would you like for authors to structure the discussion space IN THIS SITE/CATEGORY? E.g. tag comments, draw relationships between comments (repeats, counter-arguments, etc), bump up certain comments and diminish others, etc

24 (19.35%): Yes, Please!

33 (26.61%): Sounds good

31 (25.00%): Sure, why not

21 (16.94%): Maybe

15 (12.10%): No thank you

- How much would you like for the types of comments to be somehow distinguished in the interface IN THIS SITE/CATEGORY? Comments could be grouped along these lines, or at least be color coded.

16 (13.11%): Yes, Please!

41 (33.61%): Sounds good

34 (27.87%): Sure, why not

18 (14.75%): Maybe

13 (10.66%): No thank you

- How much would you like for authors to have a visual overview of their post history next to their comments IN THIS SITE/CATEGORY?

19 (15.32%): Yes, Please!

36 (29.03%): Sounds good

38 (30.65%): Sure, why not

18 (14.52%): Maybe

13 (10.48%): No thank you

- How much would you like for the relationships between commenters to be exposed IN THIS SITE/CATEGORY? E.g. how often times they've referenced each other, how often they post comments on the same page, if they're friends on Facebook, etc.

25 (20.16%): Yes, Please!

27 (21.77%): Sounds good

25 (20.16%): Sure, why not

25 (20.16%): Maybe

22 (17.74%): No thank you

- How much would you like for posters to "tag their identity" with their posts IN THIS SITE/CATEGORY? For example, one might choose to represent *chinese christian communists* for a given post. Posts would then be grouped by identity, in addition to showing the tags for a given post.

8 (6.50%): Yes, Please!

15 (12.20%): Sounds good

23 (18.70%): Sure, why not

33 (26.83%): Maybe

44 (35.77%): No thank you

- How much would you like for commenters to pose questions to other commenters in the form of a simple survey IN THIS SITE/CATEGORY? For example, in an NYTimes article on the World Cup one might ask other commenters which country they hope will win. Their answer could be instead of a freeform text comment, or integrated with it.

19 (15.45%): Yes, Please!

18 (14.63%): Sounds good

33 (26.83%): Sure, why not

32 (26.02%): Maybe

21 (17.07%): No thank you

PERCEPTION

- Do you hope to affect others' opinions/viewpoints via your comments IN THIS SITE/CATEGORY?

52 (63.41%): Yes

30 (36.59%): No

- How often do you feel there is co-operational spirit across the commenters IN THIS SITE/CATEGORY?

0 (0.00%): Always

23 (20.35%): Often

45 (39.82%): Sometimes

33 (29.20%): Rarely

3 (2.65%): Never

8 (7.08%): Not applicable

1 (0.88%): Other

- What do you think of the other commenters?

1 (0.88%): They are all idiots
16 (14.16%): Most of them are idiots, but some are ok
66 (58.41%): Mixed
27 (23.89%): I enjoy reading most comments without Schadenfreude
3 (2.65%): They're great!

■ How often do you go back to check for replies to your comments IN THIS SITE/CATEGORY?

7 (8.24%): Always
20 (23.53%): Often
16 (18.82%): Sometimes
30 (35.29%): Rarely
9 (10.59%): Never
3 (3.53%): Other

■ How important is it that your comments change others' minds IN THIS SITE/CATEGORY?

0 (0.00%): Very important
12 (23.08%): Important
16 (30.77%): Either way
20 (38.46%): Would be nice, I guess
4 (7.69%): Don't care

BASIC USAGE

■ How often do you WRITE comments IN THIS SITE/CATEGORY?

5 (4.17%): Always
18 (15.00%): Often
25 (20.83%): Sometimes
42 (35.00%): Rarely
29 (24.17%): Never
1 (0.83%): Other

■ How much time do you usually spend (per session) when WRITING comments IN THIS SITE/CATEGORY?

1 (1.11%): 60min+
0 (0.00%): 45-60min+
0 (0.00%): 30-45min
12 (13.33%): 15-30min
20 (22.22%): 7-15min
12 (13.33%): 5-7min
20 (22.22%): 3-5min
10 (11.11%): 1-3min
15 (16.67%): 0-60sec

■ How often do you READ comments IN THIS SITE/CATEGORY?

18 (15.13%): Always
57 (47.90%): Often
28 (23.53%): Sometimes
13 (10.92%): Rarely
2 (1.68%): Never

1 (0.84%): Other

- How much time do you usually spend (per session) when READING comments IN THIS SITE/CATEGORY?

1 (0.86%): 60min+

2 (1.72%): 45-60min+

9 (7.76%): 30-45min

18 (15.52%): 15-30min

20 (17.24%): 7-15min

12 (10.34%): 5-7min

24 (20.69%): 3-5min

24 (20.69%): 1-3min

6 (5.17%): 0-60sec

- What factors go into how much time you spend reading/writing comments IN THIS SITE/CATEGORY?

how bored I am

fun, DRAMA, teaching my point of view

If they are interesting (since people share ideas for projects)

How important a particular outing is to me or if I had a particularly good/bad experience

The complexity of the workout.

I only write short comments explaining why I liked something to my friends who follow my reader

How much there is to read. How much there is to write, and how much research I have to do before I can write it. I spend a lot of time correcting grammar and commenting on quirks of the English language, the structure of poems & paragraphs, etc.

Whether people will read my comment (probably not, except maybe the author of the article if the comment is early enough) and whether anyone has left an insightful comment (usually not).

Depends on what i am looking for. But generally i like to comparison shop and comments on review sites like amazon make that possible.

Also depends on whether i am looking on something i know i want or something i am wondering whether to get.

Is it relevant to something I care about or is it interesting (usually, does someone have something worth saying)

How busy I am

interest in the subject, lack of discussion

How infuriated I am by the article

level of satisfaction with product

How consistent are the comments. If there is a wide variety of opinion, then I read more. If everyone agrees, then I read less. I tend not to write product comments because I'm no longer looking at the comments on the product after I own it, so am no longer on that site or that part of the site.

I have to know the person or feel a personal connection with them.

Interest in the original story, quality of discussion. Potential to find out more information or interesting links.

If I am looking for more information.

tie strength, potential viewership/contagion/feedback

I have to know the author and feel like the feedback I provide would be helpful or at least conversationally interesting. For me, commenting on media sites is a social experience--I rarely engage in deep intellectual discourse on these sites.

Comment length

Personal relation to the person whose photo/status update I'm commenting on. Whether or not I feel like I have anything worth saying to share...

If the product was unusually bad or good - the review will take more time & be longer.

The issue being discussed

how much i want to buy the product.

Depends on the audience, what the issue/comment is about, how long it takes to think of something funny/creative to say.

how much time i have to spare

how much i need the product

how bored i am

But I don't really do this

How interesting/important the original post was.

how much time i have to spare

Reading: number of helpful reviews (pertinent information for what I'm looking at with regards to a product)

Writing: generally leave a brief, concise review, unless have a particular detail that must be explained

skimming for usefulness rather than ranting

How interesting/important the original post was.

Reading: how often I check FB (how much there is to catch up on!)

Writing: how many people's items that I want to comment on (not usually very many)

how pissed i am about the article

ie, paul carr's article on airbnb from 7/25. interesting discussion on hacker news

Interest level.

how bored i am. whether specific things are happening in friends lives.

Here's the thing about Hacker News. It's not about the structure of the comments it's about the people who are there. In my mind good commenting ecosystems are all about homophily. Paul Graham put a ton of time into cultivating the community around HN and this has created a really good base community. I am looking for an environment where the highbrow tech people have discussions about cool shit in technology and that's hacker news. if i was looking for a place where there were funny/weird photos maybe id spend more time on digg or reddit.

How interested I am in the product. If they opinions seem to differ quite a bit.

Sometimes I get bored.

My passion about the topic.

Level of interest in topic.

skimming for usefulness rather than ranting

The relevance to my own life.

How much time I have

I look for consensus, try to find the extreme points, what I should look out for, etc.

I like the idea of social media quite a bit. But for me, socializing means getting *away* from the computer. So I'm on Facebook less than once a week.

I tend to only post on controversial subjects.

When posting as myself, My agenda is to try to get people to present data and evidence, rather than arguing about opinions, beliefs, or allegiances. I will find and cite references that support my position, and perhaps some that detract from it.

When posing as my right-wing alter-egos, Wayne and Brock, I make emotionally charged and manipulative arguments, each based on a different logical fallacy.

Both approaches are time-consuming.

They go way off topic and generally end up as an inappropriate discussion that is off-topic.

whether i am in research mode for a purchase

Uncertainty about a product

The importance of what I am reviewing. The more important, the more time I will spend commenting. Additionally, if I feel that my review will affect the overall review of the item, I spend more time on it. If I am one of 300 reviews, my review will generally be shorter than if my review is one of five.

The complexity of what I'm writing about.

Depends on what is of interest, how much time I have

Paper deadlines

Am I curious what people think, how divided is the issue, etc.

How much the product interests me, how much the product costs (i.e. is it worth it?)

Procrastination.

contribute something news to the post

grammar

interest in the story I am reading

How much I want to say

My knowledge of the subject and my level of eloquence at the time.

My passion for the subject

Depends on how much clarification I think the subject of the article needs...and how ignorant various comments of others may be...

Quality, knowledge of commenters.

Content of article

I only write reviews for products that have worked well for me. Typically, I write critical or negative reviews only for films.

I want to know what's going on with people, curious to get a sense of others.

i read them more if they are humorous/interesting or provide an interesting story through the branching thread of comments starting from a single comment

I'll only read if the media in question cause me to be curious what others think.

Product in review

whether i have a strong opinion of the product and feel that my opinion would matter

where i am and what else i am doing around that time

how long its been since i last logged on, how many people have made wall postings that interest me

serendipity, just seeing another post.

something I rave about or suggestions about the media that is presented.

how much i have to say and whether or not i can be articulate in my writing

something that cheesed me off, or something I rave about. usually something I am upset about.

The apparent quality of the comments, my uncertainty about the item (when buying), the likelihood of my comment being read, the strength of my opinion

how much spare time I have

lack of personal knowledge on the item, complexity of the item, cost.

I read all comments people make on my posts / directed to me. I write little comments intended to make my friends feel good, or make them aware of something cool/interesting, or make them think well of me.

I generally speaking don't care what the general populus is thinking. I believe the few people who respond/post anything having to do with an article have any real/pertinent information to share except for their own opinions which often are based on specious information or are purely emotional.

I generally don't post any comments. If I want to communicate I send a mail. I prefer not to have everyone reading what I'm sharing with another person.

SOCIAL-CULTURAL INTERACTION

- Do you participate IN THIS SITE/CATEGORY because of its socioeconomic diversity?
13 (24.53%): Yes
40 (75.47%): No
- Do you participate IN THIS SITE/CATEGORY because of its educational diversity?
13 (29.55%): Yes
31 (70.45%): No
- Do you participate IN THIS SITE/CATEGORY because of its political diversity?

17 (45.95%): Yes

20 (54.05%): No

■ Do you care about other commenters' writing quality IN THIS SITE/CATEGORY?

25 (21.93%): No, casual & freeform is preferred

69 (60.53%): Some effort would be nice

15 (13.16%): Yes, everything should be well edited

■ How much care do you put into writing comments IN THIS SITE/CATEGORY?

22 (25.29%): Freeform -- off the top of my head and send

45 (51.72%): Small amount of proofreading

20 (22.99%): Spend time crafting comments

■ Do you participate IN THIS SITE/CATEGORY because of its geographical diversity?

19 (37.25%): Yes

32 (62.75%): No

■ What kinds of diversity from any of these categories are important IN THIS SITE/CATEGORY?

51 -- Geographical

41 -- Racial

53 -- Socioeconomic

44 -- Educational

37 -- Political

19 -- Other

6 -- I only want people similar to me

■ Do you participate IN THIS SITE/CATEGORY because of its racial diversity?

7 (17.07%): Yes

34 (82.93%): No

EXPERIENCE

■ What's wrong/missing?

Commenter's insights are usually limited, so only very few comments are worth reading.

comments do not have the same cultural norms of offline discussions (politeness, thoughtfulness, etc.)

I feel like my sage wisdom usually falls on deaf ears.

Most of them are idiots

I wish viewpoints were better organized, and I could tell what was going on.

What I read are often just messages/threads on Facebook walls. Some if it is from strangers or is irrelevant to me.

Thought.

Moderator.

I don't know why I post comments, except that its frustrating to find a product with no comments. I'd like to know the background of the reviewer. Did they give it one star because they don't know

how to use buttons, or is it really a defective product? Maybe I can tease this out from their comment style, but the whole thing is kind of a crapshoot. I'd rather just have some statistics about how many people bought the thing and how many are still using it.

No one says anything worthwhile. I don't like the news. I'm grumpy now. Harumph. This is waaay more than 15 minutes.

Depends on subject. If personal in nature, comments are important. For broader subjects like current events or other political/religious/social issues, I really don't put much credence in the opinions of the mob

Other commenters don't notice my contributions.

the format isn't great for more than humorous comments or purely social; not much intellectually stimulating conversation goes on in it

sometimes my comment got pushed too far down and no one can read it

Again, I don't believe what most people say and have little concern/regard for what the general public is thinking.

sometimes people are too boring, similar, and there isn't anything really interesting about following someone's tweets or blog postings.

Clear consensus/consensuses.

some descend into random conversations which are irrelevant or don't add new information

People who actually have experience with the product and it's rivals. I'm usually either looking for specific info or informative reviews. The other thing is that the people who most often write are either super positive or hate it and the people who hate things rarely state counter arguments well. (Where are the mediocre dislikes. Except on Yelp for restaurants. But that's a culture all its own...)

Content. Random comments are often people not adding anything or telling a story that isn't really that insightful.

Many comments are ill-informed, inarticulate, or don't add much to the discussion.

ability to agree and disagree with a view. some site have only agree buttons and not disagree

Sometimes the comments are boring.

Can't always verify authenticity of comments

Thought process.

It does not seem to flow well.

It's like throwing eggs against a brick wall.

Not enough thought put into them

i get myself in trouble a lot.

■ How often are you happy with your experience regarding comments IN THIS SITE/ CATEGORY?

2 (1.79%): Always

56 (50.00%): Often

42 (37.50%): Sometimes

11 (9.82%): Rarely

1 (0.89%): Never

0 (0.00%): Other

REPUTATION

- Briefly, why do you post anonymously IN THIS SITE/CATEGORY?

too lazy to bother with identity. if it sucks it in automagically i don't go out of my way to un-identify

I don't want my name associated to the products I bought for eternity on the internet - which google and all search engines can pick up. Do I really need my future employers and people who date me to know how I felt about a certain teeth whitener and book about plastic surgery? No.

its easy and doesn't require you to login, remember passwords...

Because they let me. I hate making accounts.

I don't want to take the time to get a persistent account, or deal with any accidental repercussions of what I say.

its easy and doesn't require you to login, remember passwords...

'Cause I'm too lazy to register usually.

- Do you care about the reputation of the accounts you use to post comments IN THIS SITE/CATEGORY?

44 (55.00%): Yes

36 (45.00%): No

- How often would you want to write something but do not because it might tarnish your *offline* reputation IN THIS SITE/CATEGORY?

0 (0.00%): Always

8 (18.60%): Often

15 (34.88%): Sometimes

13 (30.23%): Rarely

7 (16.28%): Never

0 (0.00%): Not applicable

0 (0.00%): Other

- How often would you want to write something but do not because it might tarnish your *online* reputation IN THIS SITE/CATEGORY??

0 (0.00%): Always

3 (6.98%): Often

15 (34.88%): Sometimes

11 (25.58%): Rarely

13 (30.23%): Never

1 (2.33%): Not applicable

0 (0.00%): Other

- Do you care about your reputation offline being affected by the comments you post IN THIS SITE/CATEGORY?

43 (53.75%): Yes

37 (46.25%): No

- Do you leave comments anonymously and/or with a persistent identity (login) IN THIS SITE/CATEGORY?

19 -- Anonymously

71 -- Using one user account

9 -- Using multiple user accounts

- Briefly, why do you care about the reputation of the accounts you use to post comments IN THIS SITE/CATEGORY?

I realize that facebook and other social media is not private and what I say there can be searched by potential employers etc., so I care about the reputation I present.

it is me, but i can also say ridiculous things because its not REALLY me

Because it's a community that I am a part of and it's an activity that means a lot to me in the real world.

It's me.

in case someone takes my comment seriously and wants to know who the heck I am

It's traceable.

On the one social media site I use (facebook), my identity is tied pretty closely to my real life identity.

It's all part of who I am online. I use the same handle for basically everything.

I hope my login will gain credibility and people will respect what I have to contribute

consistency and quality of my reviews reinforces my account as a reputable user and thus the reviews will be taken seriously and in the end help other users -- returning the favor of others taking the time to describe their experiences.

I hope my login will gain credibility and people will respect what I have to contribute

I use the same handle and it reflects on me.

I dont want to look like a douchebag or unfunny

Because I feel that my identity on Facebook is perceived as a digital alter ego representative of who I am in reality.

I like to be liked by other people.

so other take my views seriously

less anonymous

they are tied to offline identity

Because it affects how people I know in real life think of me.

Because my social media reputation filters into my real world reputation.

So that my reviews will mean something. The better reputation I have, the more likely someone will take my reviews into account.

In social media, one account represents me among my friends and colleagues and I limit those who can see my account to those users - so I do care about my account because it connects me directly to people in my life, unlike other user accounts I have, many of which are anonymous (i.e. don't use my real name).

to get myself known in the internet or at least have good credibility when others search for me

These are my friends (most are real friends, not virtual friends that I have never met)

Because I am an official blogger at Huffington Post and I want to maintain a level of respect for the time people put into a good blog piece.

Yelp status

It has my name in it.

it represents me (i use facebook)

because i live in a small town and sooner or later everyone figures out who did the writing

because i use the same user name for all sites and i feel connected to what I say. I don't want to be embarrassed if people find out that I'm the one posting the comments

because it's among friends and family

reflects directly on me.

because they're usually connected to my own identity.

- Do you seek to improve your reputation offline by posting comments IN THIS SITE/CATEGORY?

9 (21.43%): Yes

33 (78.57%): No

- Do you usually post anonymously IN THIS SITE/CATEGORY?

11 (64.71%): Yes

6 (35.29%): No