SANJAY KAIRAM RESEARCH STATEMENT

My research in human-computer interaction addresses implications of the large-scale digital traces left behind by our online social interactions. As we socialize to an increasing extent in computer-mediated spaces, each act of communication now leaves behind a trail of digital breadcrumbs, in the form of text, multimedia, and metadata, presenting several challenges and opportunities for both theory and design.

On one hand, the presence and persistence of these traces imposes responsibilities on designers of online social systems to help users ensure that shared content doesn't reach unintended or unwanted audiences. On the other hand, when responsibly collected, these traces can offer rich insight into patterns of human communication and power applications that leverage the collective intelligence of these newly-formed digitally-mediated communities.

My work explores these issues using an interdisciplinary lens, marrying qualitative methods with techniques from large-scale data mining and information visualization. Through this lens, I have examined social phenomena in a variety of real-world systems. Specifically, I have developed models, theory, and applications relating to the following topics: 1) modeling individual strategies for selectively sharing information, 2) exploring collective patterns of communication and information-sharing in newly-forming crowds, and 3) developing novel visualization and analytical techniques for analyzing digital social traces at scale.

Modeling Strategies for Selective Sharing

When individuals share information about themselves in offline settings, the potential audience for an act of self-disclosure is often limited to those within some physical distance. In online settings, however, a wide variety of potential audience members – who might have been separated offline by physical, social, or institutional barriers – may be collapsed into a single context (Marwick & boyd, 2011). My dissertation research explores how individuals utilize various mechanisms for audience management in order to share information selectively with different sub-groups of their contacts. My specific approach frames these issues in terms of the interplay between individual motivations or practices around sharing and features of the content being shared.

Understanding individual motivations and strategies for selective sharing. To gain insight into how individuals frame decisions around what to share and with whom, I led a study of early users of the Google+ social networking service during its launch in 2011. Google+ introduced an audience-management mechanism called *Circles*, allowing users to organize contacts intuitively into personally-managed groups. Early adopters of this system had recently organized their contacts and were actively using this tool to selectively disclose information to different audiences, providing an idealized environment for studying practices around selective-sharing and audience management.

Through a combination of surveys, interviews, and analysis of usage logs, we investigated motivations for self-disclosure and strategies for organizing and selecting audiences for pieces of shared content. Among our other findings, this work provided a clear illustration of audience management as a boundary regulation process, with individuals balancing reasons for limiting access to content – including privacy, relevance to others, and norms of acceptable conduct – against motivations to share content more widely. This research contributed to a paper presented at CHI 2012 (Kairam et al. 2012a), which has received over 100 citations and was recognized by Google as one of their "Excellent Papers of 2012".

Modeling the interplay of user preferences and content features in sharing decisions. This first study examined in detail individual strategies and considerations for sharing online, but sharing decisions are also influenced in large part by specific features of the content being shared, especially in the case of rich media such as photographs.

To address questions about how these factors interact, I led a study of photo-sharing behavior on Flickr, aimed at analyzing sharing decisions from three complementary perspectives: user preferences, photo content, and image aesthetics. This study combined rich qualitative insights from a survey of active Flickr users with quantitative analysis of metadata from 10.4M photos uploaded to Flickr. Our research identified several photography and site activity measures, including camera choices and social network features, which could help predict general user sharing preferences. In addition, using computer vision techniques, we uncovered content features corresponding to the use of different permissions settings and modeled the influence of image aesthetics on these choices.

We combined these insights into a model for predicting permissions choices for individual photos uploaded to Flickr; this model could predict permissions settings with extremely high accuracy (> 94%) for particular sub-groups of photos representing approximately 20% of our sample. In a paper to be presented at CHI 2016, we illustrate how this model could help to power applications that recommend sharing settings to users after bulk-uploading collections of photos, thus helping them target sharing decisions more effectively (Kairam et al. 2016).

Future Research. These studies illustrate how I have applied an interdisciplinary approach to real-world questions around audience management and selective self-disclosure. In my continued dissertation research, I am exploring how individuals employ a variety of mechanisms for audience management, including Groups in Flickr and multiple profiles on Tumblr, to selectively target content to different parts of their existing audience and to reach new audiences. Together, this work will develop a framework for comparing how user motivations and strategies for sharing vary in accordance with different decisions about the design of interfaces to facilitate sharing. As my career continues, I aim to take findings from this study of existing mechanisms for audience management and apply them to the design of novel mechanisms that enable safer and more satisfying experiences around sharing.

Exploring Collective Processes of Group Formation

A long line of past research in the social sciences has attempted to unravel the factors that influence individuals to associate and form larger collectives. The large-scale digital traces left behind by groups as they form online can provide valuable insight into existing theoretical questions about the processes underlying the formation of new social groups. In addition, these aggregated traces can serve as a valuable source of collective knowledge on topics of widespread interest, which could be collected and analyzed to power applications that help others.

Answering existing theoretical questions about group formation and growth. In his research on the importance of "weak ties", Granovetter hypothesized that densely-clustered communities may have difficulty growing because of reduced opportunities for recruitment (Granovetter 1973). Centola and Macy, on the other hand, have demonstrated that higher within-group clustering can help attract new members (Centola & Macy 2007). Backstrom et al.'s large-scale study of the LiveJournal and DBLP networks (Backstrom et al. 2006) essentially observes both of these phenomena – friends of group members are more likely to join if their within-group friends are densely connected, but densely-connected groups are less likely to grow overall.

My interest in reconciling these seemingly conflicting observations about group formation led to a large-scale study of groups within the Ning platform. Ning allowed users to create independent *Communities* with a variety of features, including the ability to form groups. As Ning Communities were formed independently from one another, they provided an opportunity to study groups forming in varied and separate contexts, offering a kind of micro-array for studying social aggregation processes.

This research study addressed group formation processes at two different levels. We started by exploring how individuals joined groups; we identified that distinguishing "diffusion growth", where membership spreads along existing social ties, from "non-diffusion growth" (as illustrated in Figure 1) can help to explain the complex relationship between group clustering and growth. Higher clustering *does* increase the likelihood that groups will grow through diffusion, but groups which grow primarily through diffusion reach smaller sizes overall. Thus, higher clustering encourages a certain type of growth, but this particular type is ultimately short-lived.

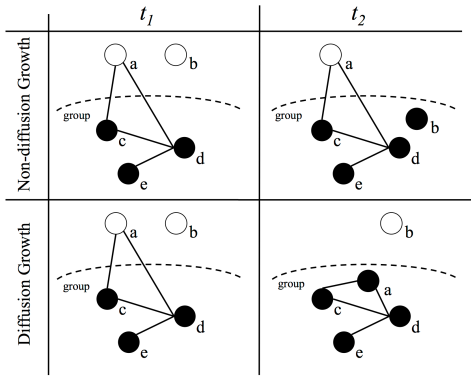

Figure 1. Illustrations of diffusion and non-diffusion growth. Differentiating between these two types of growth can help explain existing puzzles about factors influencing group formation and growth.

Using a population of 11,944 groups created in 1,713 distinct Ning Communities, we applied these insights to a model for predicting which groups would grow faster or slower. This model achieved over 79% accuracy in predicting fast-growing groups over the following two months, and 78% accuracy over the subsequent two years. This work contributed to a paper at WSDM 2012 (Kairam et al. 2012c), which has received close to 100 citations.

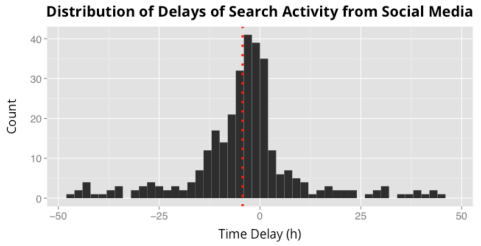

Figure 2. Our study of search and social media activity around breaking news events found that patterns of discussion on Twitter precedes Bing searches by roughly 4.3 hours, on average, providing ample opportunities for indexing social media content to help support the needs of searchers.

Leveraging the collective intelligence of ad hoc interest groups. Analyzing group formation in Ning revealed the explosive growth potential of groups that attract new members through means other than diffusion, such as associating around common interests. The internet offers many such contexts for ad hoc groups of individuals to self-assemble, such as the discussions which surface on public platforms like Twitter in the wake of breaking news events.

To explore the potential of leveraging the collective intelligence of ad hoc interest groups, we examined patterns of social media and search activity in the wake of breaking news events, as reflected through Twitter *Trends* and Bing *Trending Queries*. Combining surveys with analysis of large-scale Twitter and Bing logs, we observed differences in user search behavior related to the extent of prior knowledge about the topics involved in a breaking news event. Comparing the content shared on Twitter with search activity around these topics, we observed that social media activity tends to lead by a few hours (see Figure 2), on average, providing ample opportunity for large amounts of relevant content to be indexed and presented to users searching for information on a trending topic. Our paper on this research at ICWSM 2013 was awarded with a Best Paper Honorable Mention Award (Kairam et al. 2013).

Future Research. I am eager to continue my research on the formation of ad hoc interest groups and developing methods of mining the "data exhaust" produced by the collective interactions of these groups. Breaking news events are just one example of the types of information needs that are currently difficult to support using content produced by and indexed from traditional sources. What other types of information needs could be better supported by mining the shared knowledge of such ad hoc groups and what types of systems could help to surface this knowledge? When could the presence of newly-forming crowds be detected in order to identify content that we might want to push to individuals *before* an information need even arises?

Methods and Tools for Social Data Analysis

Many of the challenges involved in analyzing social interactions arise from the complexity of the data itself. Networks and texts both represent spaces for exploration that have high dimensionality and heterogeneity, making them challenging targets for developing effective visualization and analysis methods. My research on a variety of real-world social systems has helped to inform the design of several novel techniques for analyzing social data, which may in turn aid research on other such systems.

Providing high-level summaries of network structure. Familiar network representations, such as node-link diagrams and adjacency matrices, can be ill-suited for many high-level sensemaking tasks, such as comparing or classifying multiple networks. In response, I developed GraphPrism, a technique for visually summarizing arbitrarily large graphs through a combination of statistical diagrams, called *facets* (see Figure 3). As presented in our paper at AVI 2012 (Kairam et al. 2012b), an evaluation of this approach illustrated how even static GraphPrism diagrams could aid network analysis in high-level classification and comparison tasks with only minimal training.

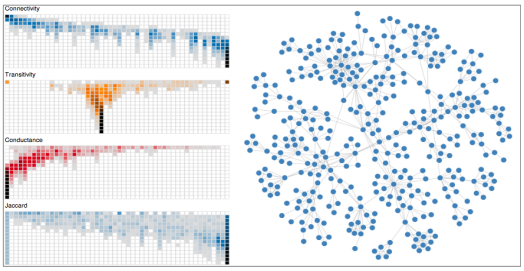

Figure 3. *GraphPrism* diagrams can summarize useful information about network clustering and connectivity that may be difficult to observe in traditional representations, such as node-link diagrams.

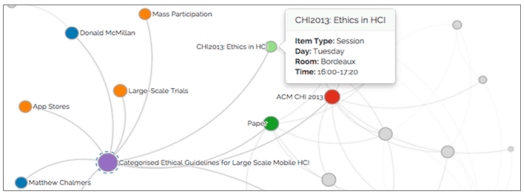

Figure 4. *Refinery* enables exploration of large, heterogeneous networks (here, conference data from *Confer*) by adapting associative strategies observed in other types of exploratory information-seeking tasks.

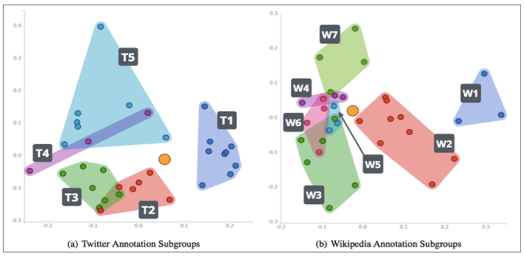

Figure 5. Annotators for two datasets, separated using *crowd parting* and plotted using PCA, with "expert labels" shown in gold. In each case, clusters do not surround the gold labels but diverge in different, but systematic ways.

Enabling bottom-up exploration of large, heterogeneous networks. Prior research in HCI has shown the value of strategies based on associative browsing, where individuals navigate from one piece of content to related items, in exploratory information-seeking tasks. Work on designing visualization systems to support these strategies, however, has remained limited. In a paper presented at EuroVis 2015 (Kairam et al. 2015), I introduced Refinery (Figure 4), a system enabling bottom-up exploration of large, heterogeneous networks through associative browsing. In this work, we identified general guidelines for designing interactive visualization systems to support associative browsing and applied these strategies to the problem of exploring complex networks. This system utilized a novel application of random-walk based graph algorithms to surface relevant content based on patches of information already explored by the user. A study in which subject matter experts used Refinery to conduct a literature review in their domain of expertise showed that they were able to quickly discover novel items of interest, including items missed in previous searches using traditional tools.

Facilitating an intermediate-level view of text annotation data. Crowdsourcing is a common strategy for collecting the "gold standard" labels required for many natural language tasks applicable to social data analysis. Crowdworkers differ in their responses for many reasons, but prior approaches have often treated disagreements as "noise" to be removed through filtering or aggregation. In a paper presented at CSCW 2016

(Kairam & Heer, 2016), we introduced a workflow design pattern that we called *crowd parting*, in which a population of workers are separated based on common patterns of responses to a shared annotation task. We illustrated this idea using an automated clustering-based method to identify diverging, but valid, worker interpretations in crowdsourced entity annotations collected over two distinct corpora – Wikipedia articles and Tweets. We demonstrated how focusing on this intermediate-level view can produce insights and power applications which are not easily gleaned from viewing either individual annotation sets or aggregated results (see Figure 5).

Future Research. By drawing on my own experience analyzing real-world social systems, I have developed novel techniques for visualizing and analyzing complex data, such as those which compose the digital traces of social interactions. In my continued research, I am eager to see how these and related approaches can be applied to the development of personal social informatics tools for everyday users of computer-mediated communication tools. By focusing on questions that system users have about the content they share, the audiences with whom they share, or the larger social context in which they are sharing, I believe we can build targeted interventions which aid these users in reasoning more carefully about the digital traces left behind by their own online social interactions.

References

[1] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, & Xiangyang Lan (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution, *KDD 2006*.

[2] Damon Centola & Michael Macy (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology 113* (3), pp. 702-734.

[3] Mark Granovetter (1973). The Strength of Weak Ties. *The American Journal of Sociology 78* (6), pp. 1360-1380.

[4] **Sanjay Kairam**, Michael J. Brzozowski, David A. Huffaker, & Ed H. Chi (2012a). Talking in Circles: Selective Sharing in Google+, *CHI 2012*.

[5] **Sanjay Kairam** & Jeffrey Heer (2016). Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks, *CSCW 2016*.

[6] **Sanjay Kairam**, Jofish Kaye, John Alexis Guerra Gómez, & David A. Shamma (2016). Snap Decisions? How Users, Content, and Aesthetics Interact to Shape Photo Sharing Behaviors, *CHI 2016*.

[7] **Sanjay Kairam**, Meredith Ringel Morris, Jaime Teevan, Dan Liebling, & Susan Dumais (2013). Towards Supporting Search over Trending Events with Social Media, *ICWSM 2013*. **Best Paper Honorable Mention.**

[8] **Sanjay Kairam**, Nathalie Henry Riche, Steven Drucker, Roland Fernandez, and Jeffrey Heer (2015). Refinery: Visual Exploration of Large, Heterogeneous Networks through Associative Browsing, *EuroVis 2015*.

[9] **Sanjay Kairam**, Diana MacLean, Manolis Savva, & Jeffrey Heer (2012b). GraphPrism: Compact Visualization of Network Structure.

[10] **Sanjay Kairam**, Dan J. Wang, & Jure Leskovec (2012c). The Life and Death of Online Groups: Predicting Group Growth and Longevity, *WSDM 2012*.

[11] Alice Marwick & danah boyd (2011). I Tweet honestly, I Tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society 13*, pp. 96-113.