Computational Perspectives on Social Phenomena at Global Scales

Jon Kleinberg
Cornell University
Ithaca NY USA

Abstract

The growth of social media and on-line social networks has opened up a set of fascinating new challenges and directions for researchers in both computing and the social sciences, and an active interface is growing between these areas. We discuss a set of basic questions that arise in the design and analysis of systems supporting on-line social interactions, focusing on two main issues: the role of network structure in the dynamics of social media sites, and the analysis of textual data as a way to study properties of on-line social interaction.

Overview

The study of on-line social interaction has provided a wide range of insights into the dynamics of social phenomena, the flow of information through the world, and the design of systems supporting the collective creation and sharing of content. Two broad recent themes in the analysis of these domains have been the use of network methods for reasoning about the patterns of interaction [Jackson, 2008; Easley and Kleinberg, 2010; Newman, 2010] and the use of text and language analysis for reasoning about the content itself [Manning *et al.*, 2008; Pang and Lee, 2008].

A promising direction is to bring these two collections of methods together, combining the analysis of what is being communicated in these systems with the structural patterns that characterize the communication. Research over the past several years has begun to explore such a synthesis for several different kinds of phenomena, including the inference of power and status in social interactions [Bramsen *et al.*, 2011; Gilbert, 2012; Otterbacher and Hemphill, 2012]; identification of different roles in these interactions [Diehl *et al.*, 2007; De Choudhury *et al.*, 2010]; analysis of how content is shaped in political contexts [Conover *et al.*, 2011; Livne *et al.*, 2011]; and investigations of how content enters news coverage and popular discourse [Leskovec *et al.*, 2009; Simmons *et al.*, 2011; Danescu-Niculescu-Mizil *et al.*, 2012a; 2012b].

Here we consider three lines of work that contribute to this combined analysis of structure and content. We begin by discussing the question of textual memes that spread on-line [Leskovec *et al.*, 2009; Simmons *et al.*, 2011], and the ways in which they penetrate everyday discourse. In particular, it is interesting to ask whether there may be specific features intrinsic to the content itself that could contribute to the success of particular memes — a kind of "fitness function" defined on the text. In order to consider this question in a controlled setting, we focus on a large corpus of movie quotes, looking for properties that distinguish lines that have emerged as memorable quotes over time from lines that have remained more obscure, despite being uttered by the same individual at approximately the same point in time [Danescu-Niculescu-Mizil *et al.*, 2012a]. We identify several features associated with memorable quotes; in particular, when evaluated on language models trained on newswire [Kučera and Francis, 1967], memorable quotes tend to use less probable word sequences, but their part-of-speech sequences are more probable. We also find that memorable quotes use features associated with greater *generality*, through choice of verb tense, personal pronouns, and articles.

For the second line of work we consider, we move from the scale of population-level content sharing down to the level of person-to-person communication: we explore the use of *language coordination* [Niederhoffer and Pennebaker, 2002] as a mechanism for identifying power relationships between people from textual traces of their interactions [Danescu-Niculescu-Mizil *et al.*, 2012b]. Language coordination is a phenomenon in which two people will tend to become more similar in their language choices when communicating with each other (for example, in the rate at which they use different types of function words such as quantifiers, conjunctions, and articles). Work in communication accommodation theory predicts that when there is a power imbalance between two people who are communicating, the lower-power person will tend to coordinate more and the higher-power person will tend to coordinate less [Natale, 1975; Giles *et al.*, 1991; Street and Giles, 1982; Giles, 2008]. Using a methodology for identifying fine-grained coordination effects in text data [Danescu-Niculescu-Mizil *et al.*, 2011], we find that coordination can serve as a useful cross-domain method for identifying power differences in both on-line and off-line domains.

Finally, we consider an application that requires an understanding of both content and structure — the automated management of on-line discussions in settings such as Facebook. Many social applications organize a user's experience around a set of discussion threads that he or she is partic-

ipating in; to capture the set of issues involved in managing this experience, we formalize the problem of *conversational curation* [Backstrom *et al.*, 2013] — at any given point in time, which threads should be brought to a user's attention? We identify two sub-problems inherent in this question. The first is to estimate the amount of discussion a thread will generate, a task related to earlier analyses of comment volume [De Choudhury *et al.*, 2009; Tsagkias *et al.*, 2009; Yano and Smith, 2010; Guerini *et al.*, 2011; Wang *et al.*, 2012; Artzi *et al.*, 2012] and rate of content diffusion [Kwak *et al.*, 2010; Lerman and Ghosh, 2010; Bakshy *et al.*, 2011; Romero *et al.*, 2011; Artzi *et al.*, 2012]. The second, which appears to be new in the present context, is *re-entry prediction* — estimating whether an individual who has already contributed to a thread is likely to continue participating. This latter task is crucial for organizing on-line conversations that evolve over time, and in particular for deciding whether to continue keeping a user informed about a thread after he or she has already participated in it.

A recurring theme in all these lines of work, and others related to them, is the way in which insights that combine properties of both content and structure can help provide individuals with richer ways of managing their interactions with information and with one another.

References

[Artzi *et al.*, 2012] Yoav Artzi, Patrick Pantel, and Michael Gamon. Predicting responses to microblog posts. In *Proceedings of NAACL (short paper)*, 2012.

[Backstrom *et al.*, 2013] Lars Backstrom, Jon M. Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proc. 6th ACM International Conference on Web Search and Data Mining*, pages 13–22, 2013.

[Bakshy *et al.*, 2011] Eitan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of WSDM*, 2011.

[Bramsen *et al.*, 2011] Philip Bramsen, Martha Escobar-Molana, Ami Patel, and Rafael Alonso. Extracting social power relationships from natural language. In *ACL HLT*, 2011.

[Conover *et al.*, 2011] M D Conover, J Ratkiewicz, M Francisco, B Goncalves, A Flammini, and F Menczer. Political polarization on Twitter. In *ICWSM*, 2011.

[Danescu-Niculescu-Mizil *et al.*, 2011] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan T. Dumais. Mark my words!: linguistic style accommodation in social media. In *Proc. 20th International World Wide Web Conference*, pages 745–754, 2011.

[Danescu-Niculescu-Mizil *et al.*, 2012a] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon M. Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, 2012.

[Danescu-Niculescu-Mizil *et al.*, 2012b] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. Echoes of power: language effects and power differences in social interaction. In *Proc. 21st International World Wide Web Conference*, pages 699–708, 2012.

[De Choudhury *et al.*, 2009] Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. What makes conversations interesting?: Themes, participants and consequences of conversations in online social media. In *Proceedings of WWW*, pages 331–340, 2009.

[De Choudhury *et al.*, 2010] Munmun De Choudhury, Winter A Mason, Jake M Hofman, and Duncan J Watts. Inferring relevant social networks from interpersonal communication. In *WWW*, pages 301–310, 2010.

[Diehl *et al.*, 2007] Christopher P. Diehl, Galileo Namata, and Lise Getoor. Relationship identification for social network discovery. In *Proceedings of the AAAI Workshop on Enhanced Messaging*, volume 22, pages 546–552, 2007.

[Easley and Kleinberg, 2010] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[Gilbert, 2012] Eric Gilbert. Phrases that signal workplace hierarchy. In *CSCW*, 2012.

[Giles *et al.*, 1991] Howard Giles, Justine Coupland, and Nikolas Coupland. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, 1991.

[Giles, 2008] Howard Giles. Communication Accommodation Theory. In *Engaging theories in interpersonal communication: Multiple perspectives*, pages 161–173. Sage Publications, Inc, 2008.

[Guerini *et al.*, 2011] Marco Guerini, Carlo Strapparava, and Gözde Özbal. Exploring text virality in social networks. In *Proceedings of ICWSM (poster)*, 2011.

[Jackson, 2008] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[Kučera and Francis, 1967] Henry Kučera and W. Nelson Francis. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

[Kwak *et al.*, 2010] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proc. of WWW*, pages 591–600, 2010.

[Lerman and Ghosh, 2010] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of ICWSM*, 2010.

[Leskovec *et al.*, 2009] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[Livne *et al.*, 2011] A Livne, M P Simmons, E Adar, and L A Adamic. The party is over here: Structure and content in the 2010 election. In *ICWSM*, 2011.

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Natale, 1975] Michael Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *J. Personality and Social Psych.*, 32(5):790 – 804, 1975.

[Newman, 2010] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[Niederhoffer and Pennebaker, 2002] Kate G. Niederhoffer and James W. Pennebaker. Linguistic style matching in social interaction. *J. Lang. and Social Psych.*, 21(4):337–360, 2002.

[Otterbacher and Hemphill, 2012] Jahna Otterbacher and Libby Hemphill. Learning the lingo? Gender, prestige and linguistic adaptation in review communities. In *Proceedings of CSCW*, 2012.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. *Opinion Mining and Sentinment Analysis*. Number 2(1-2) in Foundations and Trends in Information Retrieval. Now Publishers, 2008.

[Romero *et al.*, 2011] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. *International World Wide Web Conference*, pages 695–704, 2011.

[Simmons *et al.*, 2011] Matthew P. Simmons, Lada A. Adamic, and Eytan Adar. Memes online: Extracted, subtracted, injected, and recollected. In *ICWSM*, 2011.

[Street and Giles, 1982] Richard L. Street and Howard Giles. Speech accommodation theory. In *Social cognition and communication*. Sage Publications, 1982.

[Tsagkias *et al.*, 2009] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Predicting the volume of comments on online news stories. In *Proceedings of CIKM*, pages 1765–1768, 2009.

[Wang *et al.*, 2012] Chunyan Wang, Mao Ye, and Bernardo A. Huberman. From user comments to on-line conversations. In *Proceedings of KDD*, 2012.

[Yano and Smith, 2010] Tae Yano and Noah A. Smith. What's worthy of comment? Content and comment volume in political blogs. In *Proc. of ICWSM*, 2010.