Visual Analytics of Blog Corpora for Communication Scholars

Nicholas Diakopoulos
University of Bergen
Department of Information and Media Studies
nicholas.diakopoulos@gmail.com

OVERVIEW

We are interested in building visual analytic tools that support communication scholars in a range of analytic tasks concerning blogs. The purpose of this paper is to first describe the persona of "The Communication Scholar" as it informs our designs, and then to delve into two specific analytic tasks relating to the study of information diffusion and information polarization within topical blog networks. For each task we consider the underlying objectives of the task as well as analytics and visualizations that can support the task. We identify challenges to the design of such interfaces and hope to leave the reader with a sense for a range of interesting research questions to pursue, as well as detailed descriptions of the two tasks we describe.

Our domain of interest relates to the study of topical blog networks, in particular those that might refer to heated discussions about issues like climate change or privacy. The data we are collecting includes blogs and their posts, links, and authors, as well as links to other social media like tweets and comments. We are particularly interested in moving *beyond the analysis of blogs in purely structural terms to include text analytics techniques* such as local grammar and key statement extraction. This deep fusion of network *and* textual analysis techniques is a driving aspect of our approach and one that sets it apart from other efforts in this domain. In the next section we describe and identify some design implications for helping to analyze a blog corpus such as ours in light of our target users.

TARGET PERSONA

Personas help capture, communicate, and differentiate the distinguishing features, objectives, motivations, behaviors, and expectations of a group of intended users [1]. They are often useful as design artifacts in assessing how different design options may impact various types of users. In this section we describe a key persona which helps guide our characterization of analytic tasks in the next section.

While there are certainly many types of users who will be interested in analyzing blogs, such as media monitors, marketers, and the general public, we focus here on the persona of "The Communication Scholar" (TCS), an archetype representing a researcher interested in studying human communication as it arises in social media (e.g. blog networks in our case). In order to understand some of the objectives of TCS we draw on social science research literature and two unstructured interviews that we conducted with communication scholars who are actively engaged in research with various blog corpora. The interviews were roughly one hour each and were oriented towards eliciting relevant goals, desired features, and pain points.

Some of TCS's general goals include the exploration of data to generate or test hypotheses, deep reading of the social media content itself to understand context, and the production of illustrative material (e.g. visualizations) to communicate compelling case studies. TCS is highly motivated to spend time doing in-depth data analysis and, unlike more casual users, is willing to expend considerable energy in learning new tools and visual mappings. TCS is generally comfortable with technology but may not be able to program. More specific goals of TCS include being able to compare different subsets of data such as by group affiliation [2], link or "attention" clusters [3], genre (e.g. mommy blogs, craft blogs, main stream media blogs, other personal blogs [4]), content features (e.g. if the post contains an image), actor types [5], communicative frame [6], or language groups [7]. This goal essentially corresponds to operationalizing variables of interest to the scholar so that they can draw analytic conclusions about relationships in the data.

Other challenges include differentiating link types (such as those to other blogs, or to other blog posts), characterizing link features such as temporal durability or recency [8], and in being able to exclude material from a visual representation in order to reduce the dominance of hub nodes that might obscure other more subtle relationships. Data gathering, such as finding the initial seeds to collect blog data was also mentioned by one interviewee as a particular pain-point.

From our literature review and interviews we would argue that an important component to include in visual analytic tools designed for TCS is an ability to **create subsets of data from a blog corpus**, and ideally to **facilitate the operationalization of variables of interest** with automated or semi-automated approaches when feasible. Of particular interest here are different opportunities to use network (e.g. community detection algorithms) or textual (e.g. genre, language, or content) analysis to drive this process. Grouping data into subsets is a core sub-task of both the tasks we describe in the next section.

TASK ANALYSES

While there are many analytic tasks that TCS may engage in, here we focus on two in particular. For each we describe the core objective of the task and sub-tasks, information needs of the user, relevant analytics that support that task, and different alternatives for visual representations that support accomplishing the task objectives.

Information Diffusion

The study of information diffusion has the goal of understanding how topics, memes, or statements traverse a network over time or among different types of actors [5], [9], [10]. The core objective is often motivated by a desire to understand the dominance of, influence of, or susceptibility to different ideas. In our work we are particularly interested in how key statements, e.g. statements that characterize different positions in a debate, are taken-up in topical blog networks over time [11] .

Information diffusion is often studied with respect to the volume and rate at which an idea moves between subsets of the data such as between different types of nodes or clusters of nodes. Grouping nodes is thus an essential subtask to support, either manually through efficient interaction techniques, or automatically through analytics such as community analysis [2], language detection [12], or other facets of the data. Another aspect of grouping concerns the units of diffusion (text statements in our case) as there may be substantial similarity among these units (e.g. statements that express similar ideas), or a desire to reduce noise by aggregating units. Interfaces to support manual or semi-automated grouping of elements could, for instance, leverage visualization to indicate similarity among elements, thus making the process more efficient.

To visually represent information diffusion in abstract data we believe there could be opportunities to leverage and adapt flow depictions used in scientific visualization contexts [13]. Visually representing diffusion as flow (including the rate or acceleration of flow) between nodes on a graph may be supported by standard InfoVis depictions such as directed network representations. More traditional node-link depictions (Fig 1a) could perhaps also be creatively combined with animated flow-based visualizations such as http://hint.fm/wind/ (Figure 1b) to show a vector field of movement together with the network depiction. Techniques such as animation or small-multiples can be used to show changes in how a text statement is moving across a blog network over time.

Another analytic subtask to support is identifying interesting nodes, such as bridges or facilitator nodes between clusters or communities. Characterizing such nodes with respect to structural network properties (e.g. centrality measures, degree, use of different link types), or content properties can deepen the understanding of what makes these nodes influential or susceptible with respect to information diffusion. Standard mappings of scalar properties to color or size on a node-link network visualization should support the identification task effectively. Perhaps more challenging is finding ways to visually summarize and represent the distinctive textual characteristics of nodes through keywords or key statements in ways that support TCS's desire to be able to transition into deep reading and seeing content in context. Visualization techniques such as tag clouds, Wordles [14], and Word Trees [15] are helpful for showing overall frequency and some context of word usage but fall short in effectively supporting visual comparison.

The study of information diffusion stresses a need to simultaneously visualize three distinct information schemas (i.e. ways of organizing space): *temporal* (to understand how information moves over time), *network* (to understand what structural properties of the network facilitate or hinder information flow), and *content* (to understand what properties of content relate to flow). The challenge to visual analytic interface designers is thus to synthesize and

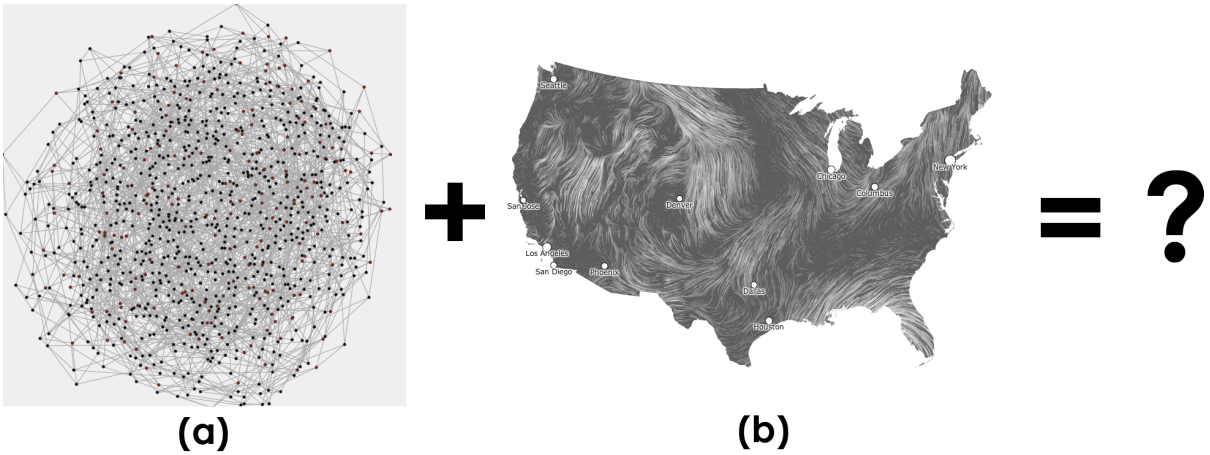

Figure 1. How might we combine traditional node-link diagrams with flow maps?

facilitate the sensemaking of interactions between these different factors (time, network, and content). An interesting interaction to help visualize would, for instance, be how a network representation changes over time based on the lifespan or recency of links between blogs. We expect visualization techniques such as brushing and linking will be essential in supporting cross-schema analyses.

Polarization

The study of polarization within information networks has the goal of understanding the degree of interaction between different groupings of the network. For example, TCS may be interested in the degree to which liberal and conservative blogs link to each other [16] or quote each other in ways that indicate agreement or disagreement. Interest in information polarization is often motivated by issues of information exposure and whether a wide array of perspectives is being considered in political debate [17]. As with information diffusion, the study of polarization is also motivated by questions of influence and information exchange between areas of a network.

Supporting TCS with capabilities to create subsets of data is perhaps an even more essential subtask for studying polarization than it is for studying information diffusion. The exact facets for creating groupings will depend on the specific research question of TCS, but many of the same facets that we identified as interesting for information diffusion would also be interesting for polarization. However, here we focus specifically on how groups of key statements may be used by some parts of the network but not others, indicating a potential lack of exchange, or biased exchange of information between those sub-networks. Thus we will define groupings based on the set of blogs and their connections that use a particular key statement.

The core task of studying polarization is in *identifying sub-networks that are poorly connected* with respect to some other unit of analysis such as political affiliation, blog genre, or as in our case use of different key statements. Identifying such networks can be aided by analytic processes such as computing the modularity of a network [12]. Modularity is an indicator of how a network's edges deviate from a reference network constructed from random edges; the higher the modularity score the less connections there are among communities within that network. Of course thresholding the modularity score would be one way to identify polarized statements, however we believe that visualization can also help facilitate discovery. For example take Figure 2, which visualizes the pair-wise modularity of hypothetical statements in a network. Darker cells immediately jump out to the user and signal pairs of potentially polarized statements (e.g. statements 9 or 10 and statement 6); somewhat darkened cells may be less polarized statements, but still interesting to consider for further investigation. Verification of polarized sub-networks, as well as more detailed characterization of the

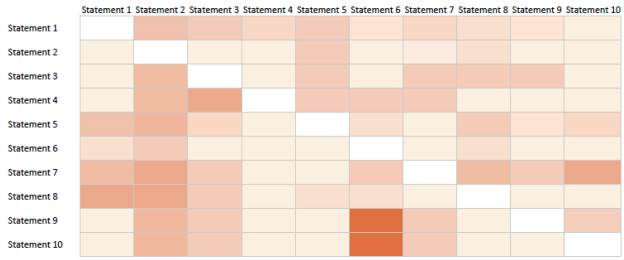

Figure 2. Matrix visualization indicating modularity among key statement networks.

degree of polarization can be supported by colored node-link visualizations of the network [11].

Another challenge to the study of polarization is in considering the semantics of network links. For instance, is a blogger linking to another blog because they agree with it, are criticizing it, providing evidence for it, or are expanding on it? Visualizing the different networks that arise from these different kinds of connections could make for some insightful comparisons. Making the intentions of these links explicit could deepen our understanding of the nuances of polarization or dimensions along which ostensibly polarized groups may actually be interacting. Thus a challenge for designing visual analytic systems for studying polarization is to develop analytical methods that can reliably extract the semantics of links in the network.

CLOSING REMARKS

In closing and to summarize, we have (1) described some of the analytic interests of a persona, The Communication Scholar based on literature and interviews, and (2) described two key tasks of interest to the persona, the study of information diffusion and polarization, including relevant analytics and visualizations that may facilitate those tasks. There are several challenges we have identified for the design of effective visual analytic tools that can support these tasks, including: (1) interfaces and algorithms to effectively group data subsets for analytic comparison, (2) synthesis of time, network, and textual analytic schemas in the same visual space, (3) development of novel text visualizations to support *comparison* among nodes in a blog network, (4) analytics to extract the semantics of links. There is much room to make advances along all of these lines and our ongoing efforts will seek to forge such paths.

ACKNOWLEDGEMENTS

This paper benefitted greatly from discussions with Dag Elgesem and Andrew Salway as part of the NTAP (Networks of Texts and People) project – http://ntap.no. This research was supported by a grant form the Norwegian Research Council's VERDIKT program.

REFERENCES

[1] D. Brown, *Communicating Design*. New Riders, 2011.

[2] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *Int. Journal of Juman Computer Studies*, vol. 65, no. 1, pp. 57–70, 2007.

[3] B. Etling, J. Kelly, R. Faris, and J. Palfrey, "Mapping the Arabic blogosphere: politics and dissent online," *New Media and Society*, vol. 12, no. 8, pp. 1225–1243, 2010.

[4] M. Hallvard, "Mapping the Norwegian Blogosphere: Methodological Challenges in Internationalizing Internet Research," *Social Science Computing Review*, vol. 29, no. 3, pp. 313–326, 2011.

[5] G. Lotan, E. Graeff, M. Ananny, M. Gaffney, I. Pearce, and D. Boyd, "The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian Revolutions," *International Journal of Communication*, vol. 5, pp. 1375–1405, 2011.

[6] A. Simon and J. Jerit, "Toward a Theory Relating Political Discourse, Media, and Public Opinion," *Journal of Communication*, vol. 57, no. 2, pp. 254–271, 2007.

[7] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. D. Bie, N. Mosdell, J. Lewis, and N. Cristianini, "The Structure of the EU Mediasphere," *PLoS ONE*, vol. 5, no. 12, 2010.

[8] A. Bruns, T. Highfield, L. Kirchhoff, and T. Nicolai, "Mapping the Australian Networked Public Sphere," *Social Science Computing Review*, vol. 29, no. 3, pp. 277–287, 2011.

[9] R. da Cunha Recuero, "Information flows and social capital in weblogs: a case study in the Brazilian blogosphere," presented at the Hypertext and Hypermedia (HT), 2008.

[10] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the Dynamics of the News Cycle," in *Conference on Knowledge Discovery and Data Mining (KDD)*.

[11] A. Salway, N. Diakopoulos, and D. Elgesem, "Visualizing Information Diffusion and Polarization with Key Statements," presented at the International Conference on Weblogs and Social Media (ICWSM) workshop on Social Media Visualization, 2012.

[12] S. Hale, "Cross-Lingual Linking in the Blogosphere," *Journal of Computer-Mediated Communication*, vol. 17, no. 2, pp. 135–151, 2012.

[13] F. Post and T. van Walsum, "Fluid Flow Visualization.," in *Focus on Scientific Visualization*, 1993.

[14] F. Viégas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, 2009.

[15] M. Wattenberg and F. Viégas, "The Word Tree, an Interactive Visual Concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, 2008.

[16] L. Adamic and N. Glance, "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," presented at the KDD Workshop on Link Discovery, 2005.

[17] C. Sunstein, *Infotopia: How many minds produce knowledge*. Oxford University Press, 2006.