Social Query: Looking for Social Signals from Online Artifacts

Jaime Montemayor and Christopher P. Diehl

Every moment, millions of people worldwide are communicating and sharing content online. On a variety of new digital social media, including discussion forums, blogs, e-mails, and status updates, we express ourselves to enrich existing relationships and to establish new relationships that would be difficult or impossible to develop offline. The staggering volume of these digital social artifacts presents new opportunities to extend and enhance our conventional notions of "search." In this article, we develop the concept of social query concept to explore the interactions of informational and social components of these new types of queries.

INTRODUCTION

Until recently, we had been accustomed to thinking about the Internet as a repository of information that, once deposited, could be retrieved by providing carefully chosen keywords to search engines. However, in the past few years, we have seen an explosion of additional types of digital artifacts that represent, explicitly or implicitly, social relationships and interactions among people. Now, more than questions such as, "how do I calculate the circumference of a circle?" we may want to ask, "I don't know how to calculate the circumference of a circle. Is there an expert who can help me?" Are standard search methods sufficient? How do we want to query for such data?

In general, even though the actual queries that we issue may be keyword based (because that is the typical input supported by current search engines), the underlying goals may contain both informational and social components. For example, "recipe for curry chicken" is a query that has only an informational element, whereas "I need to buy a new car, but I don't know what models might be good for my needs, and which dealers to visit" contains both informational elements (car models and dealers) and implicit social elements (who can help me with choosing a model and which dealers might I trust?). Notice that the informational components are aimed at identifying artifacts that potentially satisfy the under-

lying information need, whereas the social components are queries that require knowledge about the social attributes and behaviors that identify individuals, groups, or relationships exhibiting detectable and associable characteristics in the artifacts they create.

Here is another example. When we enter a person's name, "Jaime Montemayor" for example, into a search engine, we see a list of documents that in some way match the terms "Jaime" and "Montemayor." We also see a number of pages that represent Jaime Montemayor's academic and research lives. We do not immediately see references to his other (publicly searchable) interests, such as judo and Argentine tango. Of course, if we had entered "Jaime Montemayor judo" or "Jaime Montemayor tango," then the results would have contained pages that associate him with those two social activities. But this action would have required prior knowledge about his group affiliations. Indeed, when I type the terms "Jaime Montemayor," I am probably interested in more than the pages that happen to contain these two words. More likely, I want to know: Who is Jaime Montemayor? What does Jaime Montemayor like and not like? Who are friends of Jaime Montemayor? As far as we know, no existing system directly supports queries of this form.

These examples illustrate our concept of social query: query interactions with online repositories that are composed of informational and social components. We hope you agree that the queries in the example described above feel natural; that is, these are questions we ask all the time. Unfortunately we are often limited to issuing only informational queries and satisfying the social query through other more laborious means, for example, by iteratively issuing queries, reviewing the results, guessing, and refining query terms. Making the transition from our existing understanding of search to this generalized social query concept is difficult. As a first step toward a general social query capability, we focus on the subproblem of social relationship identification, the task of identifying pairs of entities that exhibit a specific social relationship (as in the friends of Jaime Montemayor described in the example above).

Along with finding these relationships, one is often interested in identifying the supporting digital social artifacts so that query results can subsequently be validated. This suggests that a comprehensive solution to the social relationship identification problem should include mechanisms for human interaction with the query algorithms. Ideally these mechanisms, or workflows, should help accelerate the discovery of relevant relationships, allow the user to validate and track hypothesized relationships, and generate reports or summaries of findings. In the rest of this article, we present two scenarios, examine the challenges posed by the task of identifying social relationships, and describe an initial realization of social relationship identification within the context of an analytic workflow.

SCENARIO 1: CONSTRUCTING A SOCIAL RELATIONSHIP GRAPH

One promising application area for social relationship identification is electronic forensic science. The e-discovery industry has emerged to provide technology to help process and identify evidence to support legal cases. Unfortunately, analysts and investigators continue to struggle with current tools available to sift through tremendous volumes of electronic artifacts. To understand the scope and complexity of the problem, one needs to look no further than the Enron scandal. Before its bankruptcy in December 2001, the Enron Corporation was one of the world's leading energy companies, with core business in the generation and distribution of electricity and natural gas. Beginning in 1998 and continuing through 2001, members of Enron devised fraudulent schemes to manipulate various energy markets for financial gain. From 2000 to 2001, these schemes were responsible for exacerbating the California energy crisis as Enron misrepresented available supply and demand. The deception ultimately led to mounting losses that could no longer be concealed, resulting in the company's stunning collapse by the end of 2001 from its peak 1 year before. During the course of the U.S. government's investigation, large numbers of documents, e-mails, and telephone calls were subpoenaed and the collection was made part of the public record, providing a rare glimpse into a large corporation through its digital artifacts. The e-mail collection in particular consists of approximately 250,000 unique e-mail messages collected from approximately 150 Enron e-mail accounts.[1] Given the complexity of the domain, the task of assembling a general picture of the events that transpired by using the e-mail data is monumental and remains daunting even with analytic tools to assist in the process.

We now know that Tim Belden, a focus of the government's investigation,[2] pled guilty to one count of conspiracy to commit wire fraud as part of a plea bargain.[3] With this information as a starting point, an investigator may want to answer questions such as "was Tim Belden acting alone or at the behest of his supervisors?" and "were his subordinates aware of his illegal activities and did they knowingly participate?" To answer these questions, one must first identify Tim Belden's relationships within the Enron corporate structure. In particular, we wish to solve two social relationship identification problems: (*i*) who are the supervisors of Tim Belden? and (*ii*) who are the subordinates of Tim Belden? In the case of Enron, ground truth exists to answer these questions; however, this will not always be the case. This is especially true in instances in which individuals engaging in criminal activity go out of their way to disguise their activities and relationships. This forensic domain provides a concrete example where identifying the sup-

porting evidence of social relationships is an important component of the relationship identification problem.

Consider the Enron e-mail communications from January 2000 through November 2001. Notice that, although the focus is on a very small subset of messages, those (from among the quarter-million messages) marking Tim Belden's communication history, the amount of data is daunting (Fig. 1).

The communications graph shown in Fig. 1 depicts all the e-mail addresses in the e-mail collection to which or from which Tim Belden sent or received a minimum of five e-mail messages during the specified time period. Additional directed edges were added among Tim Belden's contacts if this same minimum level of communication occurred. The nodes represent e-mail addresses. This representation (a common practice) provides little insight into the relationship structure surrounding Belden or the context of these relationships. Moreover, we can discern no clear group structure.

Is there an alternative process that provides an investigator with cues to relevant communications relationships along with specific e-mails that highlight a particular social relationship of interest, such as a manager–subordinate relationship? In short, yes. In prior research, Diehl et al.[4] demonstrated within the context of Enron that this is indeed possible. They introduced a machine-learning approach for learning to rank-order communications relationships and their associated messages on the basis of their relative likelihood of exhibiting a social relationship of interest. By exploiting an Enron document that specifies a series of manager–subordinate relationships that existed over the given time period, they were able to demonstrate the algorithm's ability to successfully learn to cue an investigator to relevant relationships and e-mails. In prior work, we developed an analytic workflow around this ranking paradigm, called SocialRank, which demonstrates a process for identifying social relationships in an e-mail corpus.[5] Once a relevant time period has been identified, along with an e-mail address and a social relationship of interest, SocialRank displays the communications relationships that most likely exhibit that social relation (for example, manager–subordinate) as determined by the ranking algorithm.

Figure 2 shows one of the stages of the SocialRank workflow where the user interactively explores e-mail message traffic represented as a time series. This display includes a time series for each of the top four candidates for Tim Belden's supervisor, on the basis of observed communications during the selected time period. Overlaid on each time series are visual cues that highlight particularly compelling evidence for the social relationship. In this display, for each candidate, asterisks call out the three most compelling pieces of evidence found in e-mails sent to and received from Tim Belden, and the shaded rectangular

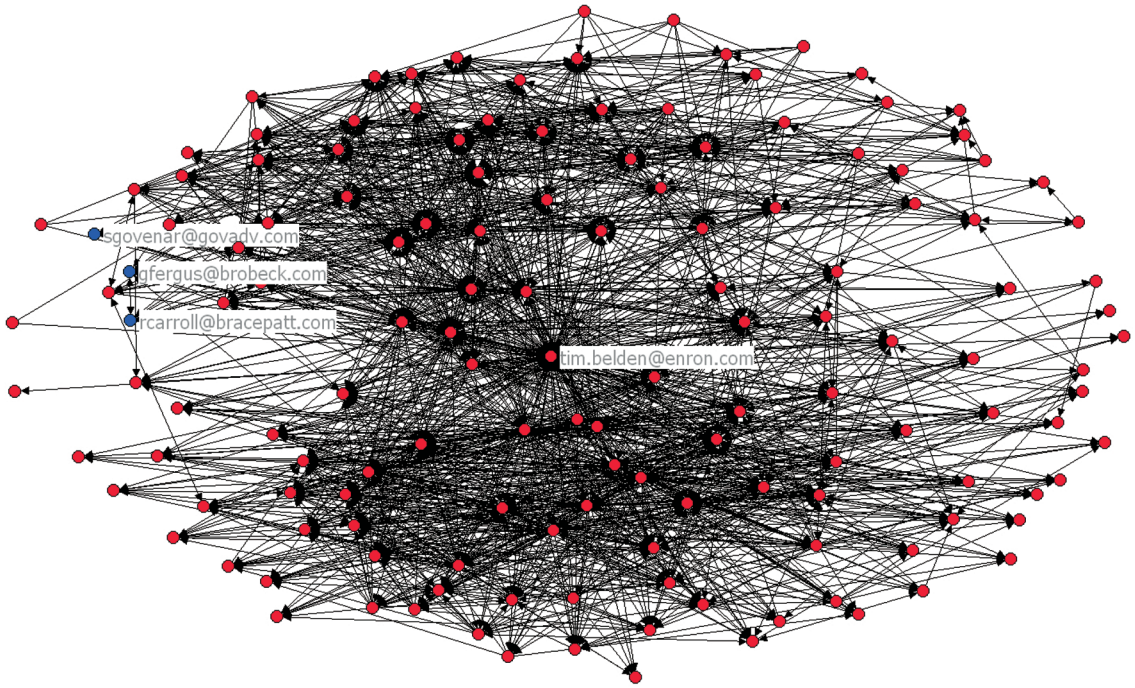

Figure 1. This communications graph depicts all of the e-mail addresses to which or from which Tim Belden sent or received at least five e-mail messages during the time period from January 2000 through November 2001. Additional directed edges were added among Tim Belden's contacts if this minimum level of communication occurred. The red nodes are Enron e-mail addresses. The labeled blue nodes are e-mail addresses outside the company. (Reprinted with permission from Ref. 6, © 2009 IEEE.)

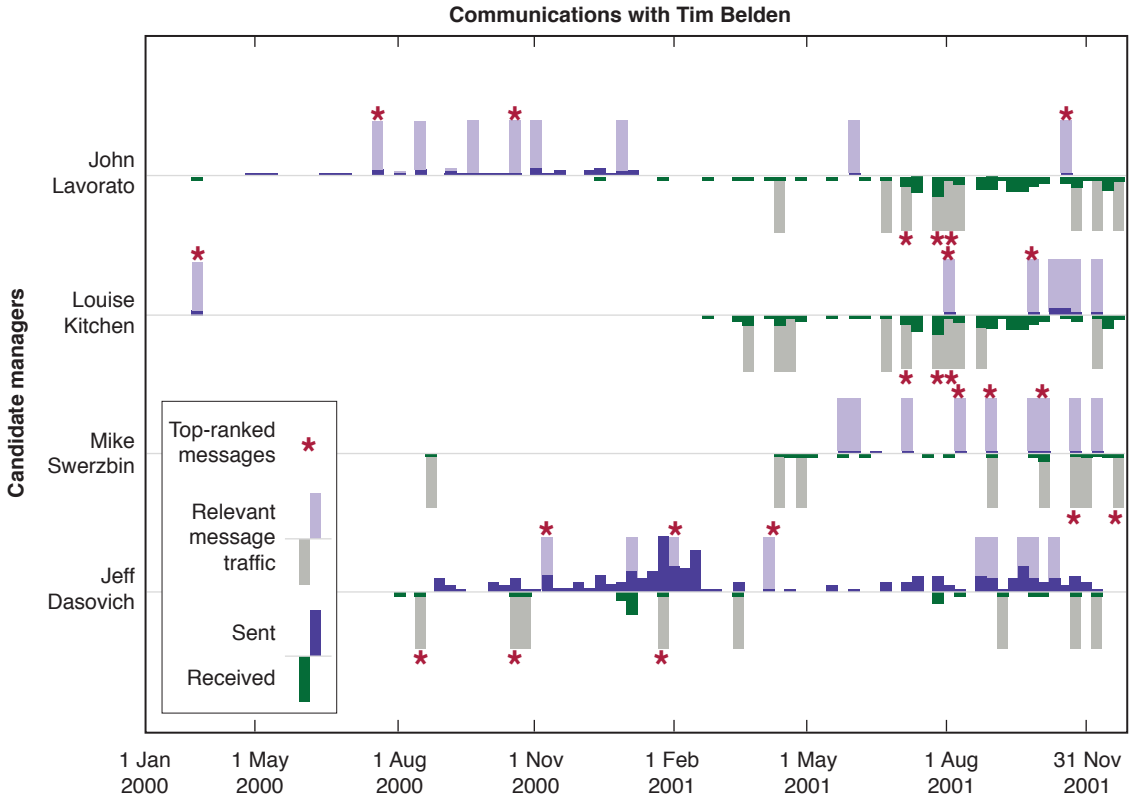

Figure 2. Examining the social query results: the e-mail relationships for the top four candidate managers of Tim Belden. Each timeline indicates the volume of e-mail traffic sent to and received from Belden weekly. The shaded bars indicate weeks with e-mail traffic supporting the existence of the social relationship. The asterisks indicate weeks with the most compelling e-mail messages between Belden and that particular candidate manager. (Adapted with permission from Ref. 6, © 2009 IEEE.)

regions indicate time intervals containing additional significant supporting evidence for the manager–subordinate relationship. These cues greatly accelerate the discovery process in two ways: First, the ranked list directs an analyst to the few most likely candidate communications histories. Second, within each history, the analyst is able to immediately focus on the most compelling entries from among the hundreds (or thousands) of messages.

Figures 3 and 4 show examples of top-ranked messages identified by the ranker and highlighted in SocialRank. The bold and italic text sections in Fig. 3 provide evidence that Belden and others reported to and received direction from John Lavorato and Louise Kitchen. The bold and italic section in Fig. 4 confirms this finding and suggests a direct or nearly direct reporting relationship. In addition to confirming these relationships, this evidence also provides a useful starting point for further elaborating the organizational structure led by Lavorato and Kitchen. Cuing the investigator directly to these messages saves time and effort by avoiding a burdensome linear search through the entire corpus. From our experience mapping the Enron manager–subordinate hierarchy,[6] we believe it would be difficult to anticipate and formulate the appropriate set of traditional keyword queries required to retrieve these messages in a comparable time frame.

SCENARIO 2: LOOKING FOR SUPPORTIVE BLOGGERS

Mary has given birth to her first child. She has been an active blogger and blog reader about her hobby, home coffee roasting (HCR). Now she is searching for personal blogs written by other women who also have just become mothers, who write with a fun and engaging style, and more important, who are supportive. From her experiences in the HCR community, she knows that devoted communities of bloggers and blog readers form when members cultivate online relationships with thoughtful interactions through posts and comments.

How can she search, using currently available technology, for similarly fun, engaging, and supportive people within the community of first-time mother bloggers and readers? One approach might be to first pose an informational query to a search engine to identify candidate "first-time mother" blogs (the underlying informational

Date: October 22, 2000
From: John Lavorato
To: Tim Belden and 9 Other Recipients
Subject: Systems

I think we are making great progress on the systems side. I would like to set a deadline of November 10th to have a plan on all North American projects (I'm ok if fundementals groups are excluded) that is signed off on by commercial, Sally's world, and Beth's world. When I say signed off I mean that I want signitures on a piece of paper that everyone is onside with the plan for each project. If you don't agree don't sign. If certain projects (ie. the gas plan) are not done yet then lay out a timeframe that the plan will be complete. I want much more in the way of specifics about objectives and timeframe.

Thanks for everyone's hard work on this.
John

Date: July 30, 2001
From: Louise Kitchen
To: Tim Belden and 55 Other Recipients
Subject: Message from John and Louise - Enron Americas Management Offsite

Please find attached details for the forthcoming Enron Americas Management Offsite. There are group actions which need to be completed before arriving in Beaver Creek. The Offsite will involve meetings, mountain biking and white water rafting (grade 3), so please bring appropriate clothing.
. . .
Video You each have been assigned to a group for the sole purpose of completing a video prior to attending the Offsite. The video filming should be completed and on a VHS tape prior to departure for Beaver Creek. The purpose of this video is to provide a comic interlude to the proceedings. The videos will be seen prior to dinner on Friday night at Saddleridge. The video should be about 5 minutes in length, on a VHS tape and there is a zero budget assigned to the production of the video. Each team has been given a title which is open to interpretation (see attached spreadsheet).
. . .
Any questions or concerns should be addressed to Dorie Hitchcock (Ext 36978) We look forward to seeing you in Beaver Creek.

John & Louise

Figure 3. Confirmatory evidence provided in top-ranked messages: John Lavorato praises his subordinates and provides additional guidance on a current task. John Lavorato and Louise Kitchen provide information and direction regarding the upcoming management offsite. (Adapted with permission from Ref. 6, © 2009 IEEE.)

Date: August 2, 2001
From: Tim Belden
To: John Lavorato, Louise Kitchen
Subject: Off-Site Travel Question

The e-mail that was sent out many weeks ago about the offsite indicated that it would run from Wednesday night to Saturday AM. It is now running Thursday until Sunday. Calger has found a leased plane that costs roughly $13k for one roundtrip and a total of $20k for two round trips. I had already made arraingements to attend a wedding in Oregon on Saturday night. It is a good friend of mine and my wife's. It's in eastern Oregon and is about a four hour drive away. I see the following choices before me:

1) Don't go to Colorado. Tell you guys that I'm a family man and not a company man.
2) Go to Colorado and fly home commercial on Friday night, leaving at about 4 PM. Incremental cost of flight would be $500.
3) Go to Colorado and fly home on the rented plane on early Saturday afternoon. Incremental cost of flight would be $7,000.
4) Don't go to wedding. Tell my wife that I'm a company man and that it is critical that I ride mountain bikes with a bunch of 30-something Enron folks all weekend.

While I have authority to place millions of dollars of the company's money at risk, I don't feel comfortable signing up for a $7,000 extra flight without talking to you guys. #3, the jet set answer costs quite a bit more, but it dramatically increases the amount of time that I spend in Colorado. #2 is cost-effective but gives me less than 24 hours in Colorado. #4, while perhaps appealing to you, doesn't work for me. #2 is probably preferable to #1, just requires a lot of travel time to me.

Any thoughts would be greatly appreciated.

Figure 4. Confirmatory evidence provided in a top-ranked message: Tim Belden asks John Lavorato and Louise Kitchen for guidance on his travel to the upcoming offsite. (Adapted with permission from Ref. 6, © 2009 IEEE.)

need may be represented by the terms "first-time mother blog"). Now, from the query results, Mary must (*i*) first scan the list to see whether any blog titles and the few lines of text interest her, then (*ii*) for each of the selected blogs, read through the posts and post comments to see whether any of the bloggers exhibit the qualities that she desires (fun, engaging, and supportive). If Mary successfully finds some qualifying bloggers, she might look for others by another approach: exploring outlinks from those few blogs. Her assumption is that the bloggers she follows have previously discovered others with a similar style and personality. These two approaches highlight the limit of current technology: we use informational and structural cues merely to identify subsets of bloggers. Mary must bear the burden of discovering, through her reading of the blog contents of first-time mother bloggers.

How might this scenario change if social query were available? Imagine a social search engine that references the publicly available digital social artifacts from Mary's current relationships with the HCR bloggers. We presume that she has been an active participant with these bloggers for some time. Therefore the digital social artifacts, such as her own blog posts, links to other blogs, and comments on other posts along with comments by other readers, reflect a rich history of interaction. This ideal social search engine would then analyze these digital social artifacts, develop an online social interaction model about Mary, and with the model, find and propose a list of first-time mother blogs that match her desires.

Unfortunately, we do not yet know how to deal with the complex issues that must be solved to realize such an

ideal social search engine. Instead, imagine the following alternative: Mary presents, to a query engine, timelines that represent her blog relationships with her favorite coffee roasting blogs. She doesn't know how to describe why she likes those blogs. But she knows what she likes. So Mary highlights the time periods that contain her favorite posts or comments. The social search engine now looks for distinguishing social signals embedded in the language and interaction styles of the bloggers.[7] These characteristics are used to rank-order potentially interesting HCR bloggers and identify particular posts that demonstrate a style of interaction similar to what Mary has experienced. Our current research involves adapting the algorithms and workflows used to solve the manager–subordinate relationship identification problem to that of finding promising blog interactions.

CONCLUSION

Social media technologies are transforming the way we connect with friends, colleagues, and strangers over time and distance. The tremendous volumes of resulting digital social artifacts can dramatically transform the way we ask questions, from "what is this gadget?" to "what is this gadget, and who might want to use it and for what purposes?" In this article, we presented the concept of merging algorithms and interactive workflows together to address these new types of social queries. We also described a case study of a specific analytic workflow, called SocialRank, which illustrates the potential utility of social queries.

ACKNOWLEDGMENTS: This work was supported by an independent research and development grant from APL.

REFERENCES

[1]Klimt, B., and Yang, Y., "The Enron Corpus: A New Dataset for E-mail Classification Research," in *Machine Learning: ECML 2004, Proc. 15th European Conf. on Machine Learning*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (eds.), Pisa, Italy, Berlin, Springer, pp. 217–226 (2004).

[2]Fergus, G., and Frizzell, J., *Status Report on Further Investigation and Analysis of EPMI Trading Strategies*, Brobeck, Phleger, and Harrison, LLP, provided by Findlaw.com at http://tinyurl.com/cqdvr6 (accessed 17 Dec 2010).

[3]*United States of America v. Timothy N. Belden: Plea Agreement No. CR 02-0313 MJJ*, United States District Court, Northern District of California, San Francisco Division, provided by Findlaw.com at http://tinyurl.com/cm3g8m (accessed 17 Dec 2010).

[4]Diehl, C. P., Namata, G. M. S., and Getoor, L., "Relationship Identification for Social Network Discovery," in *Proc. AAAI-07: 22nd Conf. on Artificial Intelligence*, Vancouver, Canada, pp. 546–552 (2007).

[5]Montemayor, J., Diehl, C., Pekala, M., and Patrone, D., "Interactive Poster—SocialRank: An Ego- and Time-Centric Workflow for Relationship Identification," in *Proc. IEEE Symp. on Visual Analytics Science and Technology (VAST 2008)*, Columbus, OH, pp. 179–180 (2008).

[6]Diehl, C. P., Montemayor, J., and Pekala, M., "Social Relationship Identification: An Example of Social Query," in *Proc. 2009 International Conf. on Computational Science and Engineering (CSE-09)*, Vancouver, Canada, pp. 381–388 (2009).

[7]Donath, J., "Signals in Social Supernets," *J. Computer-Mediated Commun.* **13**(1), 12 (2007).

The Authors

Jamie Montemayor

Christopher P. Diehl

Jaime Montemayor is a member of the senior research staff at APL's Milton S. Eisenhower Research Center. Dr. Montemayor has contributed to numerous internal and sponsored projects on a variety of topics, including human–computer interaction, information visualization, social media analysis, and synthetic world data. **Christopher P. Diehl** has been a computer scientist with the Center for Applied Scientific Computation at Lawrence Livermore National Laboratory (LLNL) since late 2009. Before joining LLNL, he was a senior research scientist at APL's Milton S. Eisenhower Research Center and an assistant research professor with The Johns Hopkins University Department of Electrical and Computer Engineering. Dr. Diehl's research interests span machine learning, natural language processing, and human–computer interaction. For further information on the work reported here, contact Jaime Montemayor. His e-mail address is jaime.montemayor@jhuapl.edu.

The *Johns Hopkins APL Technical Digest* can be accessed electronically at **www.jhuapl.edu/techdigest**.