Social Relationship Identification: An Example of Social Query

Christopher P. Diehl, Jaime Montemayor, Mike Pekala
Milton Eisenhower Research Center
The Johns Hopkins University Applied Physics Laboratory
Laurel, MD 20723
Email: Chris.Diehl@jhuapl.edu

***Abstract*—Every moment, millions of people worldwide are communicating and sharing content online. We express ourselves online to enrich existing relationships and establish new relationships that would otherwise be difficult or impossible to develop offline. Such actions are reflected in a corresponding set of digital social artifacts, such as blog posts, emails and status updates. We are accustomed to thinking of collections of digital artifacts online as repositories of information. What if we are now searching collections of digital social artifacts that reflect people's online relationships with one another and aspects of their lives? Are standard search methods sufficient? How do we want to query such data? We contend that general queries may have both informational and social components to them. We define social queries as queries about social attributes and behaviors that identify individuals, relationships or groups exhibiting such characteristics. To develop a deeper understanding of social query, we focus on the specific task of social relationship identification. In the context of two scenarios, we examine the challenges posed by the task, review an initial realization of social relationship identification and present a way forward to address the general task.**

I. Introduction

Every moment, millions of people worldwide are communicating and sharing content online. We express ourselves online to enrich existing relationships and establish new relationships that would otherwise be difficult or impossible to develop offline. As social media becomes ever more integrated into the rhythms of our daily lives, our social actions are increasingly reflected in a corresponding set of *digital social artifacts*, such as blog posts, emails and status updates. We construct, project and selectively share details of our lives with others. The staggering volume of data being generated online presents both new challenges and new opportunities.

We are accustomed to thinking of collections of digital artifacts online as repositories of information. When we have a question, we think of key words or phrases that will allow a search engine to find artifacts containing the answer. What if we are now searching collections of digital social artifacts that reflect people's online relationships with one another and aspects of their lives? Are standard search methods sufficient? How do we want to query such data?

There are at least two classes of queries a user might wish to pose: *informational queries* and *social queries*. Informational queries are queries aimed at identifying artifacts that potentially satisfy the underlying information need. Social queries are queries about social attributes and behaviors that identify individuals, relationships or groups exhibiting such characteristics in the artifacts they create.[1] General queries may have both informational and social components to them. Yet we are often limited to issuing only informational queries and satisfying the social query through other more laborious means. To clarify these points and develop a deeper understanding of social query, we begin by examining two representative scenarios.

II. Scenarios

A. Searching for Connection

Mary, an avid rock climber, is a blogger and a blog reader. She is searching for personal blogs written by fellow climbers who share her passion for the sport. She has already discovered several accomplished climbers that routinely blog in order to share their experiences. Some take time to respond to their readers through comments or blog posts that expand on the conversation. Mary, along with other blog readers, have in time developed a small but devoted community. Mary wants to find other blogging climbers located near her home that are equally engaging, fun and most importantly, supportive.

How can she search for this type of connection with other bloggers and readers using currently available technology? One approach is to first pose an informational query to a blog search engine to identify candidate blogs that are primarily geared towards climbing. From there, she must read through the blog posts and post comments to see if any of the bloggers have created the environment she is looking for. Another approach is to explore the outlinks from the blogs Mary enjoys reading. Here the hope would be that the bloggers she follows have discovered others with a similar style and personality. These two approaches highlight our limit: we use informational and structural cues merely to identify subsets of bloggers to focus on. The burden of discovering blogging climbers that are sociable and engaging lies solely with Mary.

How might this scenario change if social query was available? Imagine a social search engine that references the publicly available digital social artifacts from Mary's current

[1] In other literature, social query refers to issuing informational queries to one's social network. *Social* in our definition describes the nature of the query, not the mechanism for query execution.

relationships with the blogging climbers and fellow readers she enjoys. 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, comments on other posts along with comments by other readers, reflect a rich history of interaction. We envision a social search engine that analyzes these digital social artifacts and presents timelines characterizing the rhythms of each blog relationship. These timelines in particular highlight time periods during which the relationships are particularly active. Mary reviews these results and adjusts the suggested time periods to cover periods of the relationships that are representative of the type of connection she's seeking. The social search engine then looks for distinguishing *social signals* [1] in the language and interaction styles of the bloggers. These characteristics are used to rank order climbing bloggers and identify particular posts that demonstrate a style of interaction similar to what she has experienced.

In contrast to informational query, which has received significant attention, it appears that much less is understood about social query. We see a need for a taxonomy of social query types and their associated demands. Core challenges that we believe are universal revolve around the specification and execution of social queries. For example, in forming a social query to identify other bloggers that are compelling to Mary, some specification is needed to identify what attributes distinguish the blogging climbers she follows. In contrast to selecting keywords for an informational query, this is generally very difficult to articulate, especially in a manner that a search engine can utilize. Our natural inclination may be to describe aspects of the blogger's personality that come through their writing; yet those descriptors would somehow need to be decomposed into specifications of how language and other indicators online point to these attributes. Instead of requiring the user to formulate an explicit specification in a query, a more natural approach is to ask the reader to specify time periods during which the blogger-reader relationship is compelling. The role of the search engine is then to decipher what underlying social signals, as expressed through language and other digital social artifacts, help identify the compelling time periods.

Within this paper, we refer to this type of social query as *social relationship identification*. This task involves identifying pairs of entities that exhibit a given social relationship online along with specific digital social artifacts that support this assertion. In the next scenario, we explore a specific realization of social relationship identification within the context of an e-discovery scenario, where electronic data is explored to discover evidence in support of a civil or criminal legal case.

B. Mapping Social Relationships

When large corporations fall under investigative scrutiny, massive collections of documents, emails and other digital content are now being routinely subpoenaed to assist in the construction of legal cases. The e-discovery industry has emerged to provide technology to aid in the process of identifying relevant evidence within the volumes of data. By any account, the current tools still leave significant margin for improvement.

To understand the scope and complexity of the problem, one need look no further than the Enron scandal. Prior to 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 through 2001, members of Enron devised fraudulent schemes to manipulate various energy markets for financial gain. During the 2000-2001 time period, 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 a stunning collapse by the end of 2001 from its peak one year before.

During the course of the U.S. government's investigation, a large collection of documents, emails and telephone calls were subpoenaed and made part of the public record, providing a rare glimpse inside a large corporation through the digital artifacts they created. The email collection in particular consists of approximately 250,000 unique email messages collected from approximately 150 Enron email accounts. Given the complexity of the domain, the task of assembling a general picture of the events that transpired using the email data is monumental and remains daunting even with analytic tools to assist in the process.

As is the case with many events of this nature, we begin the investigation with some known starting points. In the case of Enron, the corporation itself undertook a review of its trading practices with the assistance of the law firm Brobeck, Phleger and Harrison LLP. The resulting memo, detailing their understanding of the various trading strategies Enron employed, helped the government focus its inquiries on those schemes and the actions of the chief trader that developed them, Tim Belden [2]. Ultimately Belden pled guilty to one count of conspiracy to commit wire fraud as part of a plea bargain [3].

Once the memo highlighting Belden's connection to the trading strategies was uncovered, a natural next step would have been to explore Belden's activities in more detail, as captured in the email evidence, in order to answer some fundamental questions: Who did Belden report to and potentially take direction from? What organizational elements was he part of? Which employees and activities did Belden supervise? When did these activities take place relative to the known events associated with the California energy crisis? In our examination, we will focus specifically on the task of identifying individuals connected to Belden that are part of the management hierarchy, either as managers or subordinates of Belden. As illustrated in figure 1, we will examine activity during the period from January 2000 through November 2001 which covers the meteoric rise of Enron during the California energy crisis along with the subsequent fall toward eventual bankruptcy.

For our first look into Belden's communications relation-

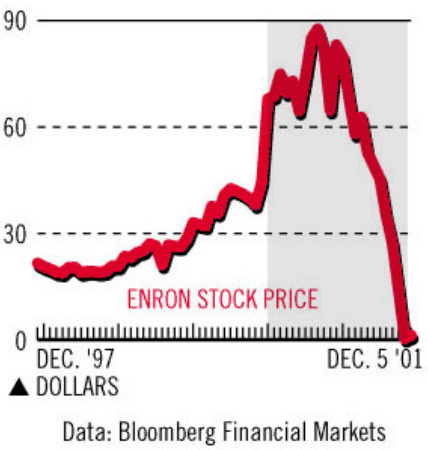

Fig. 1. The meteoric rise and fall of Enron from January 2000 through November 2001: In 2000, Enron claimed $111 billion in revenue. On December 2, 2001, Enron filed for Chapter 11 bankruptcy.

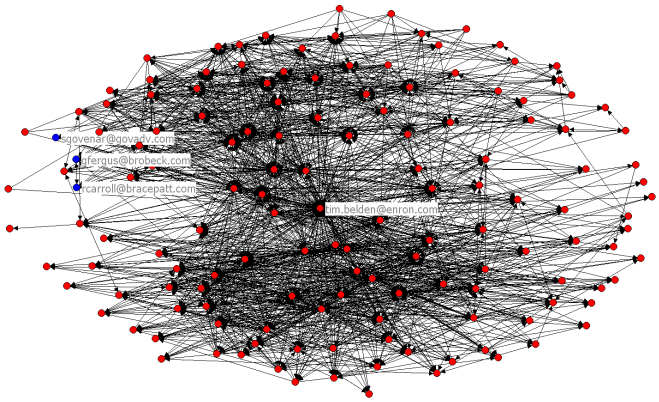

Fig. 2. This communications graph depicts all of the email addresses to/from which Tim Belden sent/received a minimum of five email messages during the time period from January 2000 through November 2001. Additional directed edges were added among Tim Belden's contacts if this minimum level of communication occurred. The red nodes are Enron email addresses. The labeled blue nodes are email addresses outside the company.

ships, consider the communications graph shown in figure 2. This communications graph depicts all of the email addresses in the email collection to/from which Tim Belden sent/received a minimum of five email messages during the specified time period. Additional directed edges were added among Tim Belden's contacts if this minimum level of communication occurred. The red nodes are Enron email addresses. The labeled blue nodes are email addresses outside the company. One of the three blue nodes is the email address for Gary Fergus, an attorney at Brobeck who contributed to the previously referenced memo detailing the investigation into Enron's trading strategies [2].

As one can clearly see, this representation provides little insight into the relationship structure surrounding Belden. At best, one might have hoped to see some indicators of group membership in this representation; yet no clear group structure emerges. The graph also provides no insight into the nature of

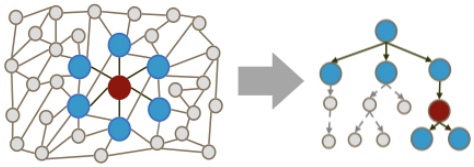

Fig. 3. Egocentric social relationship identification: rank ordering the communications relationships and email messages associated with a given email address (ego) in order to aid the investigator in discovering the underlying social relationships.

these communications relationships. A single communications relationship spanning the specified time period may exhibit indications of multiple social relationship types and the evolution through different stages.

So the fundamental question is: can we derive a process that cues an investigator to relevant communications relationships along with specific emails that highlight a particular social relationship of interest, such as a manager-subordinate relationship? In prior research, Diehl et al. have demonstrated within the context of Enron that this is indeed possible [4]. They introduced a machine learning approach for learning to rank order communications relationships and their associated messages based on their relative likelihood of exhibiting the social relationship of interest. Exploiting an Enron document that specifies a series of manager-subordinate relationships that existed over the given time period [5], they were able to demonstrate the algorithm's ability to successfully learn to cue an investigator to relevant relationships and emails. As depicted in figure 3, their process assumes that an investigator will focus incrementally on communications relationships associated with a single email address of interest, otherwise known as the email address' ego network, as she navigates the communications graph.

In prior work [6], we have developed an analytic workflow around this ranking paradigm called SocialRank that demonstrates a process for social relationship identification in an email corpus. Once the relevant time period has been identified along with an email address and social relationship ranker of interest, SocialRank displays the top ranked communications relationships that most likely exhibit the specified social relation along with cues to particular time periods with compelling message traffic. Figure 4 illustrates the top four candidate manager's communication relationships with Belden over the given time period with cues to relevant message traffic. The shaded intervals indicate weeks with email traffic supporting the existence of the social relationship. The triangles indicate weeks with one of the top three most compelling email messages.

The top two candidate managers, John Lavorato and Louise Kitchen, were the CEO and COO of Enron Americas respectively. From Belden's plea agreement [3], we know that he initially held the position of Director of Enron's California energy trading desk followed by Vice President and Managing Director in charge of Enron's West Power Trading Division in Portland, Oregon. What we do not know is when this

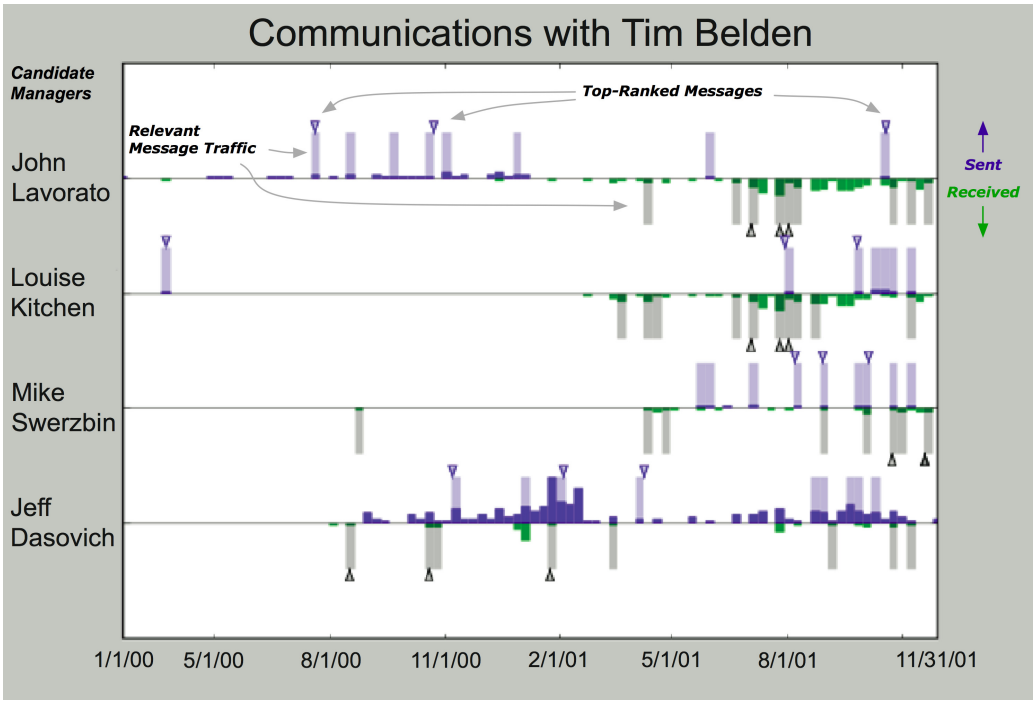

Fig. 4. Examining the social query results: The email relationships for the top four candidate managers of Tim Belden are displayed. Each timeline displays the volumes of email traffic sent to and received from Belden on a weekly basis. The shaded intervals indicate weeks with email traffic supporting the existence of the social relationship. The triangles indicate weeks with one of the top three most compelling email messages.

role change occurred. The timelines suggest that Belden began communicating to both Lavorato and Kitchen regularly through 2001 which may provide some indication of when he was promoted. Given that both Lavorato's and Kitchen's email folders are part of the email collection, the communications relationships should be fully observable. Yet there are signs that some email is missing; so caution is still warranted.

Figures 5 and 6 provide examples of the top-ranked messages highlighted by the ranker for review. The messages clarify that Belden and others report to and receive direction from John Lavorato and Louise Kitchen. Each message individually provides clear evidence of a manager-subordinate relationship. The broadcast messages from Lavorato and Kitchen also provide immediate paths for additional exploration to identify the structure of the organization that Lavorato and Kitchen lead.

Through the cuing provided by the ranker, the investigator is saved significant time and effort by avoiding the burden of a linear search through the messages. From our experience mapping the manager-subordinate relationships shown in figure 7, we believe it would be difficult to anticipate and compose appropriate queries to retrieve many of the suggested messages. The ranker therefore gives the investigator rapid exposure to a wider range of evidence without the difficulty of composing an explicit query.

III. Social Relationship Identification

When describing social attributes and relations, we naturally think in terms of adjectives that capture the elements of what we've experienced. We may indicate for example that a particular person is humorous or a relationship is supportive, with each adjective representing a host of characteristics. For a search engine to understand what constitutes a humorous person or a supportive relationship, it must understand what social signals led a user to those conclusions. Often times we struggle to articulate the indicators that we integrate so seamlessly into an interpretation. Yet when presented with examples that possess these attributes, we have relatively little difficulty identifying them as such.

In domains such as the blogosphere, where users create, share and communicate about their digital social artifacts, a natural question is whether we can leverage the perspectives of others to aid social query. Collaborative filtering (CF) allows one to discover potentially relevant content through other users that demonstrate similar preferences. To leverage CF for social query, we would need to be able to identify other users that categorize social attributes and relations in similar ways. If digital social artifacts were annotated with social metadata by the crowd, CF might provide significant value for social query, provided that coverage of the relevant digital social artifacts is sufficient.

As an example, consider a variation of the social bookmarking site *del.icio.us*. Instead of bookmarking URLs and tagging them with topical descriptors, imagine bookmarking bloggers that one finds compelling and tagging them with adjectives that capture their personality and style. As the user is composing a new entry, the site provides recommended tags for the blogger from the *social folksonomy* that emerges. Over

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

Fig. 5. 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.

time, recommendations are provided about other bloggers to consider that are both topically and socially relevant, based on discoveries of other users that share similar preferences.

Unfortunately such a rich collection of social metadata created by the crowd is not readily available today. Even if social metadata did exist in open environments geared toward sharing content, it is not clear that the coverage would be sufficient to provide utility in the long tail. We believe that regardless of whether social metadata is available, personalization based on the direct interpretation of the underlying social signals will provide benefit. This is especially true in cases such as the e-discovery scenario where social metadata is not anticipated given the nature of the communication.

To realize social relationship identification and other forms of social query, we require a system that allows the user to provide relevance cues in a way that is natural, through examples, leaving the search engine to discern the attributes

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 off-site 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.

Fig. 6. 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.

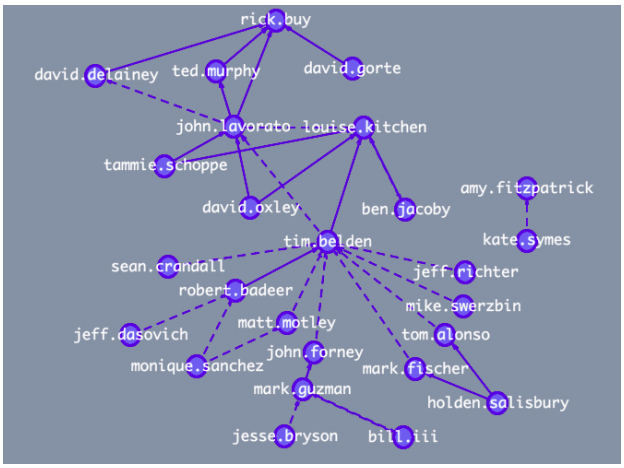

Fig. 7. Validated manager-subordinate relationships discovered using SocialRank over three days (24 hours) of exploration. Dashed edges represent relationships we believe are more than likely to be true. Solid edges represent relationships we believe to be true with high confidence.

that indicate relevance. In the following, we consider how this type of social query impacts the overall retrieval process. Broadly speaking for any type of retrieval process, there are three general steps: *query specification*, *retrieval* and *relevance assessment*. We consider each of these in turn.

A. Query Specification

We envision a general query for social relationship identification with three components:

1) Relevant and Irrelevant Relationships: First, a set of relevant online relationships are highlighted by the user. This involves identifying the online relationships that exhibit the social relationship of interest and the time periods within each online relationship during which the social relationship is present. In the blog scenario, Mary would highlight her own relationships with the other climbing bloggers that matter to her along with the time frames in the past when the relationships have been particularly enriching. In the e-discovery scenario, the relevant relationships were derived directly from the Enron document defining a host of manager-subordinate relationships for a given time period [5]; yet relevant relationships could also be specified from initial exploration of the data. At the onset or following a previous query, the user may also choose to specify irrelevant relationships in order to improve the query.

The task of explicitly specifying the time periods may seem potentially onerous. Ideally we would like the search engine to derive relevant stages directly from the digital social artifacts associated with our online relationships. In some instances, this may be possible if the relationship has been particularly active. However, the absence of artifacts does not necessarily imply the lack of relevance. Therefore we believe it is important to have the capability to make explicit assertions of relevant time periods. This provides the user with the flexibility to be clear when the available signals are not necessarily self-evident. In cases such as the e-discovery scenario, we believe explicit specification will often be far less burdensome than search with only informational query. New interactive visualizations will be needed to support this task, especially in cases such as the blog scenario where one is navigating heterogeneous collections of digital social artifacts. In the blog scenario specifically, this will be important to remind a blog reader about the evolution of their relationships with various bloggers.

2) Candidate Selectors: Second, the user specifies a set of attributes that a given relationship must satisfy in order to be considered a candidate. This may include selectors such as a topical specification consisting of key words or phrases or a structural specification describing how the candidate relationships must relate to other entities. This provides the user with explicit control over the candidate selection process.

3) Query Time Period: Finally, the user may want to define a query time period within which they are looking for social relationships of the specified type. In the context of the e-discovery scenario, this allows an investigator to search for social relationships around the time of a particular event. Note that the query time period may often be distinct from the time periods specified in the definition of the (ir)relevant online relationships. For example, in the context of the blog scenario, Mary may highlight a past relationship with a blogger as the type of social relationship she would like to find with other bloggers now.

B. Retrieval

1) Subset Selection: Once the query is specified, we want the system to generate a ranked set of online relationships along with specific references to digital social artifacts that highlight the nature of each relationship. The first step toward this goal is to retrieve the training and candidate sets of online relationships. The *training set* corresponds to the set of relevant and irrelevant relationships defined by the user in the query. Each relationship is a set of digital social artifacts created by the entities involved in the relationship over the specified time period. The *candidate set* corresponds to the set of online relationships spanning the time period of interest that satisfy the specified selectors.

In the blog scenario, the training set consists of the blog entries, blog comments and links shared among the pair of entities in each relationship highlighted by Mary. Note that unlike offline relationships, it is common to have relationships online that are asymmetric. The blogger may not be aware of the blog reader; yet there can be an active blogger-reader relationship. The candidate set consists of the blog entries, blog comments and links shared among the candidate climbing bloggers and their readers.

In the email scenario, the training set consists of the emails exchanged between the entities involved in the relevant and irrelevant communication relationships. The candidate set consists of the emails exchanged in the communication relationships associated with the ego networks of interest.

2) Learning to Rank: With the training set specified, the next step is to learn a ranker to prioritize online relationships and highlight relevant digital social artifacts. The approach taken by Diehl et al. can be described as a type of *multiple-instance preference learning*. *Instance preference learning* involves learning a scoring function $f(x)$ that maximizes ranking performance with respect to a given measure over a set of pairwise rank constraints

$$f(x_r) > f(x_i) \; \forall \; x_r \in \mathcal{X}_R, x_i \in \mathcal{X}_I \tag{1}$$

that assert the relevant instances (examples) should score higher than the irrelevant instances [7]. This is for all possible pairings from the sets $\mathcal{X}_R$ and $\mathcal{X}_I$ which are the sets of relevant and irrelevant examples respectively. Multiple-instance preference learning is a relaxation of instance preference learning where sets of examples are labeled as relevant or irrelevant as opposed to individual examples [8]. In the task of social relationship identification, the sets of examples are the sets of digital social artifacts corresponding to the relevant and irrelevant relationships. In issuing the query, the user has specified relationships, and therefore sets of digital social artifacts, that are relevant or irrelevant; yet the user has not asserted which digital social artifacts establish (ir)relevance. Our goal is to let the learning process uncover which artifacts matter.

In the e-discovery scenario, the digital social artifacts are all of a single type: email messages. The collection of filtered email messages corresponding to a communication relationship is summarized by a high-dimensional term frequency

vector that is simply the summation of the individual term frequency vectors for each email. Ranking the messages within a given communications relationship is accomplished by ordering them based on each message's contribution to the overall relationship score. When using a linear model, the decomposition of the overall score is trivial as shown in [4].

In the blog scenario, matters become more complex. A set of digital social artifacts associated with a blog relationship will often involve artifacts of different types such as blog posts, post comments and post links. Therefore relational learning algorithms that can handle heterogeneous, linked data will be important [9]. We believe discriminating social signals exist in both the link structure and the content, as is evident in the previous experiments conducted by Diehl et al. [4]. The task of constructing and integrating features from these dimensions will be a challenge.

Assessing the relative contributions of individual artifacts may require a fundamentally different approach as well in the relational context. In fact, one may argue that linking patterns alone provide significant indicators of certain relationship types in the blog scenario. This implies that individual artifacts may not be as compelling as subsets of linked artifacts that capture an interaction over time. For example, a blog post that leads to a series of exchanges in the comments. There is a tradeoff that must be addressed as we attempt to model more complex underlying patterns. More complexity requires more examples to avoid overfitting the data. This burden may not correspond with the realities of how many relationships the user is willing to identify.

C. Relevance Assessment

Once the ranker is learned from the training data, it is applied to the candidate set to rank order the relationships along with the digital social artifacts. In contrast to the e-discovery scenario, where individual email messages are highlighted, the challenge of presenting the evidence to highlight a candidate blog relationship's relevance is more complex. Irrespective of the ranking task, there is a fundamental question of how one should effectively display the history of a blog relationship for review. The cues provided by the ranker to artifacts or collections thereof will need to be layered upon this visualization. The intelligibility of the cues will be a function of the features chosen and the visualization paradigm. Research will be needed to determine a synergistic combination that aids the user in rapidly verifying relevance.

IV. Discussion

The motivation for social query clearly rests on the supposition that rich, meaningful social connection is possible online. This notion is one that has been challenged mainly on the basis that computer-mediated communication denies one access to rich nonverbal cues which provide significant evidence for making social judgments. Studies have made clear that in fact people do routinely form substantive social ties online and that relationship formation is supported by their ability to communicate social signals in other ways through the online medium [10], [11], [12], [13], [14], [15], [16], [17]. In some research, specific text-based features have been identified that aid in discriminating between relationships with various emotional states and levels of trust [15], [16], [17]. Ethnographic studies such as [12], [13] highlight that some online communities providing enriching and supportive social environments are mainly distinct from the offline social networks of the community members. Occasionally the strength of the online relationship leads to the development of offline relations as well.

Research on methods for analysis of the underlying social signals is developing. At this point, emphasis appears to be mainly on exploiting signals in content. Gill et al [18], [19] recently investigated the ability of human raters and content analysis algorithms to identify various emotional states from short blog texts. The content analysis approaches leveraged the Linguistic Inquiry and Word Count (LIWC) content analysis tool and Latent Semantic Analysis (LSA). Human raters were able to successfully identify emotions such as joy, disgust, anger and anticipation. Fear was most successfully identified by the LSA technique, exceeding the performance of the human raters. Oberlander and Nowson [20] and Nowson and Oberlander [21] investigate author personality classification from blog text, demonstrating promising results for identifying four important personality traits.

For analysis tasks that focus on individuals, content analysis alone may be sufficient. When focusing on relationships, examining multiple artifact types may be required. Gleave et al [22] advocate for this approach when analyzing social roles online. Without examining both structural and contextual clues, they claim it would have been impossible to understand the nature of the roles. The qualitative nature of their analysis emphasizes the degree of the challenge posed by heterogeneous artifact data. Currently one must search for and articulate explicitly a pattern associated with the observed behavior. We want to move beyond this to a system where the user can simply make high level relevance assessments, leaving the system to identify the underlying social signals that are discriminative.

To date, discussion of the challenges and opportunities presented by social media search appears limited. A notable exception is the paper by Hearst et al [23] that examines the unique challenges posed by blog search. For the specific task of identifying blogs / authors to read, they identify the need to specify both informational and social attributes in the query. They recommend a faceted search interface to accomplish this task. Within this interface, social dimensions would appear that describe different style and personality characteristics of bloggers. Classifiers would be trained to map bloggers to these specific dimensions.

What is unclear about the envisioned system proposed by Hearst et al is whether or not personalization is supported. The authors seem to suggest pre-trained classifiers would automatically identify bloggers with various style and personality characteristics (e.g. "witty, snarky, serious, empathetic"). As discussed earlier, we believe the interpretation of social

attributes is quite personal. Therefore we are advocating for a retrieval process that leverages relevant and irrelevant examples to learn what underlying social signals define the relationships of interest to the user.

V. Conclusion

Social media technologies are transforming the way we connect with friends, colleagues and strangers over time and distance. Researchers are beginning to understand that while such channels of communication may appear shallow relative to offline communication, people adapt to the medium and form meaningful social connections. Social query facilitates connection by helping people find others online that are creating artifacts they find engaging.

To develop a deeper understanding of social query, we examined the specific task of social relationship identification. In the context of two scenarios, we explored the challenges posed by the task, reviewed an initial realization of social relationship identification and presented a way forward to address the general task. Future work will focus on defining the taxonomy of social query types and understanding their associated demands.

Acknowledgment

This work was supported by an internal research and development grant from JHU/APL.

References

[1] J. Donath, "Signals in social supernets," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, 2007, http://jcmc.indiana.edu/vol13/issue1/donath.html.

[2] G. Fergus and J. Frizzell, "Status report on further investigation and analysis of EPMI trading strategies," Brobeck, Phleger and Harrison, LLP, http://tinyurl.com/cqdvr6.

[3] "United States of America v. Timothy N. Belden: Plea agreement," http://tinyurl.com/cm3g8m.

[4] C. P. Diehl, G. M. S. Namata, and L. Getoor, "Relationship identification for social network discovery," in *AAAI '07: Proceedings of the 22nd National Conference on Artificial Intelligence*, July 2007.

[5] "Response to request no. 11: Identify each person that designed, valued, marketed, executed, or hedged energy forward contracts, swaps, and options maturing or requiring payment anytime from January 1, 2003 through December 31, 2006." http://tinyurl.com/cfsooc.

[6] J. Montemayor, C. Diehl, M. Pekala, and D. Patrone, "Poster: SocialRank: An ego- and time-centric workflow for relationship identification," in *IEEE Symposium on Visual Analytics Science and Technology*, 2008, http://tinyurl.com/67uwqg.

[7] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 137–144.

[8] C. Bergeron, J. Zaretzki, C. Breneman, and K. P. Bennett, "Multiple instance ranking," in *ICML '08: Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 48–55.

[9] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*. The MIT Press, November 2007.

[10] J. B. Walther, "Interpersonal effects in computer-mediated interaction: A relational perspective," *Communication Research*, vol. 19, no. 1, pp. 52–90, 1992.

[11] J. J. Preece and K. Ghozati, "Observations and explorations of empathy online," in *The Internet and Health Communication: Experience and Expectations*, 2001, pp. 237–260.

[12] M. Whitty and J. Gavin, "Age/sex/location: Uncovering the social cues in the development of online relationships," *CyberPsychology and Behavior*, vol. 4, pp. 623–630, 2001.

[13] N. F. Ali-Hasan and L. A. Adamic, "Expressing social relationships on the blog through links and comments," in *ICWSM '07: Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[14] U. Pfeil and P. Zaphiris, "Patterns of empathy in online communication," in *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 919–928.

[15] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 929–932.

[16] J. T. Hancock, K. Gee, K. Ciaccio, and J. M.-H. Lin, "I'm sad you're sad: Emotional contagion in CMC," in *CSCW '08: Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*, 2008, pp. 295–298.

[17] L. E. Scissors, A. J. Gill, and D. Gergle, "Linguistic mimicry and trust in text-based CMC," in *CSCW '08: Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*, 2008, pp. 277–280.

[18] A. J. Gill, D. Gergle, R. M. French, and J. Oberlander, "Emotion rating from short blog texts," in *CHI '08: Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 1121–1124.

[19] A. J. Gill, R. M. French, D. Gergle, and J. Oberlander, "The language of emotion in short blog texts," in *CSCW '08: Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*, 2008, pp. 299–302.

[20] J. Oberlander and S. Nowson, "Whose thumb is it anyway?: Classifying author personality from weblog text," in *Proceedings of the COLING/ACL - Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 627–634.

[21] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large scale personality classification of personal weblogs," in *ICWSM '07: Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[22] E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith, "A conceptual and operational definition of 'social role' in online community," *Hawaii International Conference on System Sciences*, vol. 0, pp. 1–11, 2009.

[23] M. A. Hearst, M. Hurst, and S. T. Dumais, "What should blog search look like?" in *SSM '08: Proceedings of the 2008 ACM Workshop on Search in Social Media*, 2008, pp. 95–98.