Interactive Poster - SocialRank: An Ego- and Time-centric Workflow for Relationship Identification

Jaime Montemayor* Chris Diehl† Mike Pekala‡ David Patrone§

Milton Eisenhower Research Center, The Johns Hopkins University Applied Physics Laboratory

ABSTRACT

From instant messaging and email to wikis and blogs, millions of individuals are generating content that reflects their relationships with others in the world, both online and offline. Since communication artifacts are recordings of life events, we can gain insights into the social attributes and structures of the people within this communication history. In this paper, we describe SocialRank, an ego- and time-centric workflow for identifying social relationships in an email corpus. This workflow includes four high-level tasks: discovery, validation, annotation and dissemination. SocialRank combines relationship ranking algorithms with timeline, social network diagram, and multidimensional scaling visualization techniques to support these tasks.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User interfaces; H.4.3 [Information Systems]: Communications Applications—Information browsers; I.3.6 [Computing Methodologies]: Methodology and Techniques—Interaction techniques

1 INTRODUCTION

From instant messaging and email to wikis and blogs, millions of individuals are generating content that reflects their relationships with others in the world, both online and offline. As networked groups and organizations increasingly leverage online means of communication and collaboration, there is an opportunity to develop insights regarding the structure, attributes and dynamics of the underlying social network from such data.

Before we can understand the actions of a networked organization, it is important first to understand the social structure that supports the organization in its desire to achieve particular strategic objectives. Once the structure is sufficiently well understood, we can then focus on organizational dynamics and place particular actions of organizational elements in perspective. As individuals within the organization act toward realizing their intentions, their actions will often generate artifacts, leaving traces of these events. Clearly not all events will be observable. Yet as more communications shift to the online domain, a plethora of digital artifacts will remain.

In intelligence analysis and litigation support, where an analyst needs to reconstruct a representation of the social network from the data with minimal context, the process involves mapping the communications graph, which represents communication events among network references (email addresses, telephone numbers, etc.), to a validated social network expressing typed relationships among the known entities that the analyst believes are substantiated by the data. Within this process, there are two distinct tasks: entity resolution and relationship identification. In this paper, we focus on the relationship identification problem: the identification of relevant communications that express a specified social relationship. We first highlight the algorithmic components that support discovery and validation. Then we review how these components are integrated in SocialRank with visualization and interaction methods to facilitate annotation and dissemination, thus completing the analytic workflow.

*e-mail: Jaime.Montemayor@jhuapl.edu

†e-mail: Chris.Diehl@jhuapl.edu

‡e-mail: Mike.Pekala@jhuapl.edu

§e-mail: David.Patrone@jhuapl.edu

2 RELATIONSHIP IDENTIFICATION

Informal, online communications are composed of structured and unstructured data. At the most basic level, this includes the network references corresponding to the sender and one or more recipients, the date and time of the communication and the message content. We define a communications archive as a set of observed messages exchanged among a set of network references. Every archive has a corresponding communications graph that represents the message data as a set of dyadic (pairwise) communication relationships among the network references. The task of relationship identification involves identifying a mapping from the dyadic communications relationships to one or more social relationships of interest. In this section, we discuss two classes of algorithms that support the analyst in the construction of the underlying social network: content-based and activity-based relationship ranking.

2.1 Content-Based Relationship and Message Ranking

We envision an analyst navigating the communications graph by following and incrementally investigating ego networks. We use a two-step process to identify relevant social relationships (e.g. manager-subordinate) within a given ego network. Using a scoring function learned from message content associated with labeled ego networks [1], communications relationships are first ranked according to their relative likelihood of exhibiting a specified social relation; then, the messages within each communications relationship are ranked according to their relative support for the relationship rank.

2.2 Activity-Based Relationship Ranking

Once we have identified a particular social relation of interest, we often want to discover other communications relationships that may indicate the existence of group structure within which the identified social relationship is embedded. We achieve this by comparing the patterns of communication between a given reference communications relationship and the remaining relationships within the ego network. This provides a purely structural approach that helps the analyst establish relationship similarity, independent of content, thereby complementing the content-based rankers learned from analyst annotations.

Given a collection of activity vectors that represent the temporal rhythms of the relationships in the ego network, we use metric multidimensional scaling to generate a two-dimensional configuration of points that represents the relative similarities of the relationships, as captured by the Euclidean distance among the original activity vectors in the high-dimensional vector space. By selecting

a particular communications relationship to serve as the reference, the remaining relationships can be resorted based on their distance from the reference.

3 SocialRank

The utility of a workflow for relationship identification is dependent on its ability to 1) dramatically accelerate the discovery of relevant relationships, 2) validate and track hypothesized relationships, and 3) generate reports to disseminate an analyst's findings. SocialRank facilitates discovery and validation through a combination of ranking algorithms and information visualization techniques. An analyst discovers interesting relationships using the timeline (Figure 1), multi-dimensional scaling (MDS), network structure, network evolution (Figure 3), and message viewers. The timeline viewer displays an ego's pairwise communication relationships over time. The content-based relationship ranker orders the communications relationships in terms of their relative likelihood of exhibiting a user-specified social relationship (e.g. manager-subordinate). In order to assert that such a social relationship exists between an ego and alter, the analyst inspects the communications relationship timelines of the candidate alters. Since the most important messages supporting the relationship are indicated with visual cues on the timeline, instead of wading through hundreds of email messages, the analyst is directed to a few messages to read in detail to assess whether the content supports the relationship. Hence, this combination of relationship ranking and visualization can accelerate the discovery of messages containing supporting evidence.

When a message supports a social relationship, an analyst asserts this claim and creates an annotation. SocialRank then automatically inserts the new validated relationship into the network structure diagram (Figure 1), and remembers the corresponding email message and notes (Figure 2).

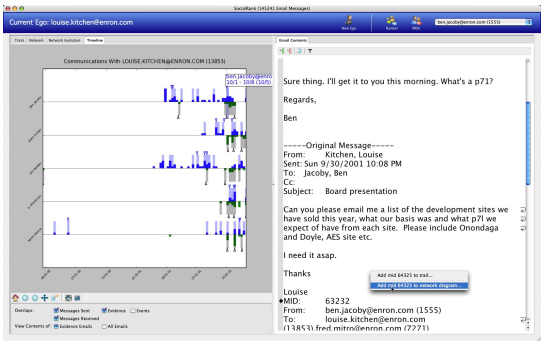

Figure 1: The timeline viewer displays an ego's pairwise communication relationships on a timeline. A supervisor ranking algorithm highlights (light shading and triangles) the time intervals that contain messages that are likely to express this relationship. After reading a message, if an analyst is satisfied that the content suggests a social relationship exists between the ego and alter, s/he can immediately create an annotated relationship (through a contextual menu) and assign the message as the validating evidence.

The MDS diagram complements the timeline. It relies on a structural comparison between the reference and candidate communications relationship over a specified time interval. Thus, once a reference ego-alter pair has been identified with a social relationship, an analyst can use the MDS diagram to reveal additional candidates by examining other communication pairs that exhibit similar patterns relative to the reference.

The network diagram in Figure 2 represents the captured knowledge of social relationships, their corresponding validating messages and the analyst's annotations. This static diagram cannot represent relationships that develop and terminate in time. We developed the network evolution viewer to incorporate the temporal attribute (Figure 3). In this diagram, SocialRank tracks the evolution of a social network (centered on an ego) and shows the temporal locations of the messages (evidence) that support the relationship.

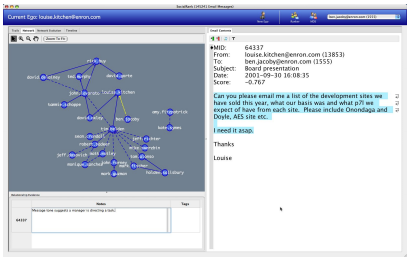

Figure 2: SocialRank automatically tracks an analyst's discoveries about social relationships and their corroborating email messages and annotations.

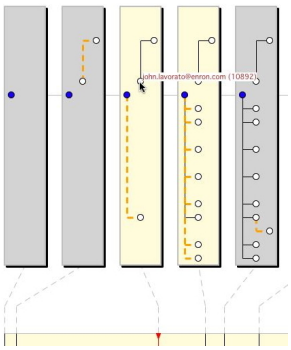

Figure 3: The egocentric network evolution viewer shows an analyst's development of the egocentric organizational structure and the temporal locations of the supporting evidence.

Finally, SocialRank demonstrates a complete workflow by supporting the dissemination (reporting) phase of the analytical process. When an analyst is ready to present the results of her work, SocialRank will collect the data of entities who are connected by a social relationship to an ego and generate an HTML-based report, including a summary network diagram, followed by the evidence and comments on each relationship in that network.

4 Next Steps

Our next machine learning objective is to develop and integrate an incremental learning capability into SocialRank so that rankers can be incrementally trained as the analyst provides annotations during exploration of the communications archive. The current content-based ranker implemented in SocialRank was trained offline on manager-subordinate relations in the Enron email corpus derived from Enron documentation. To move toward automated incremental learning, a series of additional challenges must be addressed such as learning from partially labeled ego networks with uncertainty in the time extent of the social relationship and automated model and feature selection. Such methods will be integrated with new information visualization techniques to better represent time in both the network evolution and MDS views.

References

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