A methodology for integrating network theory and topic modeling and its application to innovation diffusion

Jana Diesner and Kathleen M. Carley, Carnegie Mellon University
School of Computer Science
Institute for Software Research
Pittsburgh, PA, USA
diesner@cs.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract— **Text data pertaining to socio-technical networks often are analyzed separately from relational data, or are reduced to the fact and strength of the flow of information between nodes. Disregarding the content of text data for network analysis can limit our understanding of the effects of language use in networks. We present a computational and interdisciplinary methodology that addresses this limitation by combining theory from socio-linguistics with social network analysis and machine learning based text mining: we use network analysis to identify groups of individuals who assume the theoretically grounded roles of change agents and preservation agents. People in these roles differ in their motivation and capability to induce and adopt change in a network. Topic modeling is then constrained to the texts authored by people in these roles. We apply this methodology to a public dataset of about 55,000 research proposals that were granted funding. Our results suggest that the people per role differ in the research domains they work on and the strength of association with those domains that both roles are involved with, but are similar with respect to fulfilling the task or additional role of being a project manager.**

Keywords- topic modeling; unsupervised learning; socio-technical networks; research funding

I. INTRODUCTION

Network data comprise nodes representing the entities in a socio-technical system, edges representing the interactions between the nodes, and additional information that help to contextualize and interpret the graphs that are formed from the nodes and edges [1]. We refer to graphs as relational data. When network data are collected, natural language text data pertaining to the networks can often be acquired along the way. These text data can help to transform relational data into network data. An example for the joint availability of relational data and text data are emails, where the social network constructed from the headers can be enriched with information from the bodies [2]. Another example are social media data, where people are not only linked through pointers, but also through the content and comments they provide [3]. However, text data and relational data are often either analyzed separately, or are reduced to the fact and frequency of the transmission of data or objects between nodes [4]. This approach disregards the substance of information and communication. It also limits the insights that can be gained from performing network analysis for cases in which considering text data and network data leads to a more comprehensive understanding of socio-technical networks than using only either one data type [5, 6]. We address this limitation by contributing a computational and interdisciplinary methodology that utilizes relational data and text data by:

- Using network analysis to identify the set of social entities who assume structural roles that are defined in a theory developed and evaluated in socio-linguistics [7].
- Constraining topic modeling - an unsupervised machine learning technique that reduces the dimensionality of text data to the gist of a body of information - to the identified entities [8].

This methodology can be applied to questions, domains and datasets other than those used herein. To further facilitate the assessment of this approach we use publically available data and software. This paper focuses on methodology rather than on analyzing a socio-technical network. Therefore, the results are of illustrative nature rather than of empirical importance. Zooming out from the specific use case presented here we envision this methodology being one useful strategy for tackling problems that evolve around the following question: What topical profiles are used by individuals and groups who assume theoretically-grounded roles in networks that make them prone to actuate or inhibit change and innovation within socio-technical networks?

II. BACKGROUND

A. Network-Centric Theory of Permanence and Change

Based on empirical analyses of the diffusion of vernacular through the social network of people living in Belfast, Ireland, Milroy and Milroy [7] identified the following relationships between the roles of individuals and their disposition and ability to motivate or impede language change in the wider network: first, innovators are marginal to any adopting group, weakly linked to various groups, geographically and socially highly mobile, and under-conforming to deviant Second, early adopters of innovations are central and strongly tied members of the adopting group. Their impact on diffusion is easier to measure than that of innovators. Third, members of close-knit, dense and multiplex networks benefit from the group's intrinsic capacity to provide support and assistance and to resists external pressures, such as modifications of group-specific values, norms and standards, but can also be constrained by the group. In summary, the diffusion of innovation requires the

This research was supported in part by ARI W91WAW07C0063, GMU MURI FA9550-05-1-0388, and ONR MMV N00014-06-1-0104. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Institute, the Office of Naval Research, or the United States Government.

absence or dissolution of a connection between network structure and the members' behavior regarding the innovation.

Milroys' model of the actuation of language change draws from Granovetter's theory of strong and weak ties [9]. This theory postulates that people typically acquire novel information not from their direct contacts, but through ties that reach far into the network. Milroys' model is consistent with diffusion theories originating from other fields [10] and with theories that assume people from the middle of the social hierarchy to spread innovation rather than people from the lowest or highest social strata [11]. The model however conflicts with theories that assume opinion leaders such as mass media, subject matter experts, and celebrities to be most effective in diffusing innovations [12]. Finally, the network features that the Milroys identified were more powerful in explaining language change than alternative extra-linguistics factors such as status, class and socio-demographics.

B. Methodological Background and Contribution

We leverage Milroys' theory about the relationship between network roles and language change as a strategy for jointly considering relational data and text data. In the following we refer to innovators and early adopters as described above as *change agents*, and to other members of tight cliques as *preservation agents*. The methods section details how we operationalize these roles by using network analysis. In order to identify the language use associated with these roles, a text mining method is needed that scales up to the size of real-world corpora, can be applied to different domains and text sets, and identifies the themes addressed by each role. Topic modeling lends itself to this task: the method results in a user-defined number of unlabeled words clusters called topics. Each topic consists of text terms, and each term has a probabilistic weight that indicates the strength of association of a word with a topic [8]. The identification of topics is a non-exhaustive and non-exclusive process, i.e. not all terms belong to a cluster with other terms, while some terms are highly indicative for multiple topics. Finally, topic modeling is an unsupervised machine learning a method. Therefore, neither a sizeable ground truth is needed, which is expensive on many dimensions to obtain, nor any a priori defined patterns or features.

Using network analysis to select the entities on which topic modeling will be performed is a simpler approach than alternative solutions that add variables to latent structures. The philosophy underlying topic modeling assumes these structures, which are presented as a probabilistic graphical model, to have probabilistically generated the observed text data. Topic modeling is then used to infer this structure from the data by using Bayesian inference. Our approach draws from two streams of prior extensions to topic modeling: Chang et al. [13] took topic modeling to the network text analysis level: they perform topic modeling on the bag of words that surround the pairs of nodes in text data that form links in order to suggest link labels. Mimno and McCallum [14] have formalized the addition of variables to probabilistic graphical model: they argue that not only topics and documents have probability distributions over topics, but also other types of meta-data data, such authors and dates. The adopted the vanilla version of Latent Dirichlet Allocation (LDA) [15], and modifying it into the Dirichlet-Multinominal Regression (DMR) method. DMR takes probability distributions of various types of meta-data into consideration. DMR differs from the methodology presented herein as follows: first, our approach abstracts away from the level of individual authors to the structural role level. These roles are grounded in network theory that originates from socio-linguistics. Second, our methodology is a two-step process that links network analysis to topic modeling as opposed to a one-step process that is described by a graphical model in which roles are a variable. Consequently, with our methodology, topic modeling is performed on individuals and aggregates of individuals who assume roles that are defined and detected independently from the topic modeling process. These roles furthermore are one mechanism that drives the behavior of real-world, large-scale, socio-technical networks. With our approach, an individual can assume multiple roles, e.g. wearing their change agent's hat on one team and their preservation agent's hat on another. While we demonstrate our methodology for two specific roles, other people might be interested in analyzing the language use of (aggregates of) people assuming other roles. The presented methodology lends itself to such adoption and customization.

III. DATA

Federal funding agencies who administer tax payers' dollars increasingly release publically accessible information about the allocation of money to people and ideas. This trend has increased the transparency over state-level decision making processes for everyone. Furthermore, it has provided large and rich data sources that are suitable for investigating various questions about socio-technical networks, e.g. on collaboration and the stimulation and diffusion of innovation in certain regions and domains, and graph analytical questions such as the properties and behavior of large-scale, real-world networks.

For this project, we collected the circa 60,000 proposals that the European Union (EU) accepted for funding under the "Framework Programmes for Research and Technological Development", short Framework Programmes (FP). The data are publically available through the Community Research and Development Information Service (CORDIS) [16]. The FPs were established by Research Council of the EU in 1984 and have been continued and funded them since then; with the 7th FP being underway. The overarching goal with the FPs is to stimulate and enable competitive research in the European Research Area (ERA). For each project, CORDIS provides the full name, affiliation, and contact information for the project coordinator; a role equivalent to the principal investigator in the US. The same information is given for each collaborator on a project. For each project, CORDIS also specifies the start and end date, costs and amount of funding awarded, completion status, various keywords and index terms, and three fields of unstructured, natural language text data. These data contain the title, description ("objective"), and additional information ("general information") per project. The cumulative length of the text data varies widely across projects; ranging from short summaries to verbose descriptions of the background, methodology and technical details per project. The National Science Foundation (NSF), the US American equivalent to the EU Research Council, also provides public data on the projects they funded, which allows for future comparisons of findings.

The completeness of records in CORDIS varies. Table 1 provides descriptive statistics on the data per FP. In this paper, a "project" means a CORDIS database entry for which at least a unique identification number is provided. There are 55,972 projects for FP1 to FP6. A "project with text" means that the objective plus general information for a single project exceed a length of ninety characters. We established this heuristic to eliminate headers of empty slots in the project template, such as "research objectives and content". A "project with person" means that for at least one person on the project, a non-empty and valid entry is given in the name field. Valid entries exclude descriptors, such as "not available", "Address", and "TBC".

TABLE I. DATASET STATISTICS

FP	time range	number of projects	ratio of projects with text	ratio of projects with person	ratio projects with text and person	valid entry for person	ratio of distinct people
1	1984–1987	3283	82.7%	78.3%	71.1%	3246	88.1%
2	1987–1991	3884	79.9%	63.0%	57.9%	8545	86.9%
3	1991–1994	5529	76.8%	65.7%	60.8%	18411	87.1%
4	1994–1998	15061	79.9%	93.1%	74.5%	58692	84.1%
5	1998–2002	17629	75.4%	95.6%	72.3%	75355	59.3%
6	2002–2006	10255	96.9%	90.4%	87.4%	46201	82.2%

One of the biggest challenges with this dataset was locating the various instances and spellings of people's names so that they can be mapped to one consistent name per actual individual. High accuracy in this step is crucial because errors during the normalization of names, which are then converted into nodes of a networks, get propagated to the structural level, where they cause biases in network structure and respective analytical results [17]. In order to identify the various ways by which people are referred to we developed a data-driven set of rules and heuristics: first, all gender and role identifiers, such as "Mrs." and "Professor" were removed from the names. Single-letter umlauts were converted into their diphthong equivalent. All tuples of identically spelled names were considered to represent one and the same person if their institutional affiliation and/or their address matched completely or in at least three consecutive tokens, where tokens are any combination of space separated letters and/or digits. The word "the" was disregarded in this process. The last row in Table 1 shows how many of the name mentions per FP (second-last column) are unique individuals. People for who no valid name entry was given were ignored since they could not be disambiguated or mapped to other instances of their persona.

For this project, we normalized the data for FP1 to FP6 only because projects accepted under FP 7 are still being added such that the data for FP7 might be incomplete. For further analysis, we focus on FP4 to FP6, because these programs are significantly larger and more complete than FP1 to FP3.

IV. METHOD

The methodology presented herein is a two-step procedure that combines network analysis with topic modeling. In order to operationalizing the theoretically grounded roles of change agents and preservation agents we use network analytic metrics and respective values that are indicative of these roles. This strategy requires the construction of graphs to begin with: for each FP, we built one graph by linking the project coordinator to every collaborator on a given project. Collaborators were not linked to each other in order to avoid overly dense clusters that might not reflect the reality of collaboration on research grants. Therefore, our network formation approach results in a star structure as opposed to complete cliques per project. Stars are networks where nodes link to one central node, but not to each other beyond that. Multiple instances of pairs of collaborating people are reflected in the cumulative edge weight.

Table 2 shows an overview on basic network statistics: About 30% to 43% of the nodes are isolates; i.e. nodes that have no tie to any other node. Here, isolates represent people who are the only recipient of a grant, or who are the only person in a multi-person project for who a valid name entry was specified. Another large portion of nodes are pendants, i.e. peripheral nodes that are linked to only one other node. Also, the majority of links are connections to pendants. A large number of pendants can be connected to one and the same individual; resulting in a marginalized power structure that may exhibit norm enforcing behavior. Also, pendants are unlikely to motivate innovation because their betweenness centrality score equals zero. More than 99% of the links in each graph have a weight of one; indicating that most pairs of people collaborate on no more than one project during a FP. The density of each graph, i.e. the number of actual connections over possible connections, is very low (below 0.00005); suggesting very sparse networks. This characteristic is typical for large socio-technical networks. According to the data, not only the number of links per node and the length of text per project follow a skewed distribution, but also the amount of funding awarded per project: in FP6, for example, one percent of the projects was given 24.6% of the total amount of funding (total of over 17.5 billion Euros, which is about over 22 billion US Dollars).

TABLE II. NETWORK STATISTICS

FP	Number of nodes	Ratio of isolates	Ratio of pendants
4	49343	29.6%	44.3%
5	44675	42.8%	56.3%
6	46201	31.8%	42.9%
FP	**Number of edges**	**Ratio of edges with weight > 1**	**Ratio of edges to pendants**
4	31532	0.1%	84.4%
5	26974	0.4%	69.2%
6	31362	0.8%	83.0%

There is no canonical set of metrics and respective values or ranges of values that represents the roles of change agents and preservation agents. There is also no such set for the concepts of strong and weak ties, which are closely related to these roles. We identified the set of measures that seemed best suited to capture the notion of the roles of interest by reviewing the network analysis literature and properties of network analysis measures. Table 3 lists the measures that we selected (column 1), a short definition of each measure and selected value ranges in square brackets in column 2 (for full definitions of these measures see [18]), and an abstract value per measure and role in columns three and four. We used the ORA software

for network analysis [18]. We removed the isolates from the graphs because they do not impact the selected measures, but they do slow down computations. Next, we computed the measures listed in Table 3 for each node. We ranked the nodes per graph by their scores across all the measures considered. Since the scaled values for centralities were very small, we worked with the unscaled values. We then evaluated the top and bottom per measure and chose cut off points where values started to increase or decrease steeply. Since there are no predefined interpretations of the values of most network measures, such as 0.8 to 1.0 = high, we determined the cut offs based on the data as detailed in the squared brackets in Table 3. We established and followed these guidelines for each graph representing FP4 to FP6 based on the data. This decision-making process is not fully automated, but requires data-driven and case-wise decisions. Thus, the presented methodology requires a basic understanding of network analysis.

TABLE III. OPERATIONALIZATION OF AGENT ROLES*

Measure	**Definition and meaning of measure**	**Change agents**	**Preservation agent**
Centrality measures:			
Degree	Sum of direct links (undirected) [0 to 2 = low, 3 to 9 = medium, 10+ = high]	medium	high
Betweenness	across all node pairs with shortest path containing node *i*, how many paths pass through *i* [0 = low, 1 to 9 = medium, 10+ = high]	high	low to medium
Closeness	Average distance of a node from all other nodes. [equally low for all nodes]	high	low to medium
Centrality related measures:			
Potential boundary spanner	High in betweenness centrality AND low in degree centrality [betweenness/ degree >1.25 = high; degree – betweenness >= 20 = low (since for these cases betweenness often = 0 division not meaningful]	high	low
Actual boundary spanner	Cut point, binary (their removal results in a new component)	likely	unlikely
Triad based measures:			
Triad Count	Number of triads (closed triangles) centered at a node [0 = low, 1 = medium, >1 = high]	low	low to high
Cluster based measures:			
Clustering Coefficient	Density of subgraph induced by a node's ego network (immediate neighbor) [0 = low, 0.1to 0.5 = medium, >0.5 = high]	low	high
Clique Count	Number of distinct cliques to which a node belongs where clique = maximal complete subgraph of 3 or more nodes [0 = low, 1 = medium, >1 = high]	low	low to high

*[values based on empirical data]

For each node that we identified to represent a change agent or perseveration agent we retrieved all text documents including the title, objective and general information for each project that these people are associated with. Table 4 details how many people were identified for either group per FP, how many projects these people are associated with, and for how many of these projects there is text data available that exceed a cumulative length of ninety characters. Based on our role operationalization there are more people who assume the role of preservation agents than change agents in this dataset.

TABLE IV. FREQUENCY OF INDIVIDUALS AND TEXTS PER ROLE

Role	Metric	FP4	FP5	FP6
Change agents	Number	66	164	48
	Projects	291	1720	225
	Projects with text	251	1547	200
Preservation agents	Number	107	176	306
	Projects	158	1611	388
	Projects with text	123	1384	375

In the next step, we used Mallet [19], an open-source software product, to perform topic modeling on the retrieved text sets. After various pretests we decided to generate 20 topics per text set regardless of variations in text set size. This decision is based on our observation that the dominating topics and associated key words kept showing up in a fairly robust fashion across experiments with 5, 10, 20 and higher numbers of topics. We chose to use the LDA approach in Mallet as opposed to the topical N-gram modeling in order to avoid an over-fitting to dominating content-domain specific phrases.

V. RESULTS

Performing network analysis and topic modeling follows a rigid procedure and explicit computations. In contrast to that, interpreting the resulting topics is a fairly non-standardized process that leaves ample room for reading meaning into the results [20]. While we have some expertise in a few research domains and also some experience in acquiring funding for research, we are not qualified to evaluate topic models for grants awarded over the last 16 years and across a broad range of domains. Thus, our interpretation of the results serves to illustrate the usage of the method rather than providing empirical insight. A thorough, domain oriented analysis of the results would require the involvement of subject matter experts.

Table 5 shows the 12 most likely topics per role for FP6 sorted from left to right by decreasing values of the Dirichlet parameter (DP) and the 9 most likely terms per topic sorted by decreasing strength of association with a topic. The Dirichlet parameter indicates the strength of likelihood of a topic among the retrieved topics (the higher the more prevalent). We chose the topic labels based on the terms and term rankings per topic. Incoherent topics were disregarded. The results suggest that for both roles, "project management" is the dominating topic, which might represent a general task that leaders of any type have to handle. Preservation agents load stronger on this topic than change agents do, and also the 2nd most prevalent topic for preservation agents centers on general terms related to research in the EU. For change agents, the 2nd topic features the terms

"training" and "networking", which are also observed for preservation agents, but with a lower weight. For change agents, a third of the shown topics represent research related to the environment, namely waste management, energy, pollution and emission. Preservation agents address this issue in two topics. Topic mainly associated with change agents only are regional development and engineering. For preservation agents, these topics are cancer research and genetics, which might represent resource-intense fields that are more costly than other areas, industry in the context of manufacturing, and nuclear energy, which is not displayed since the respective DP is only 0.039 (16^{th} ranking topic), but represents a research area that was central when the FPs were started in the 1980ies. Terms found for only either one role among the 20 highest ranking topics are "excellence" for preservation agents, and "innovation" for change agents.

In FP4 and FP5, the dominating topics also center on project management, and lower ranking topics address specific content domains. As already observed for FP6, these topics occasionally overlap in subject matter, but differ in prevalence. In both, FP4 and FP5, change agents are strongly associated with research related to the environment. In FP6, change agents focused even stronger on this area, but preservation also addressed this topic. In FP4 and FP5, preservation agents show a focus on transportation and related industries.

VI. Conclusions, Limitations, and Future Work

We have presented a computational and interdisciplinary methodology for integrating theory from socio-linguistics with network analysis and machine learning based text mining. The resulting methodology enables users to jointly consider relational data and text data when analyzing large-scale, socio-technical networks. By using this approach, multiple types of behavioral data, namely interactions between people in the form of collaboration as well as language use, can be taken into account for analysis. This approach considers the substance of text data and helps to integrate different aspects that drive the properties and dynamics of networks.

Our findings from applying the proposed methodology to a dataset about research funding suggest that the roles of change agents, i.e. individuals who are likely to stimulate change and innovation in a given system, and preservation agents, i.e. people in the position to establish and enforce standards and norms, relate to certain differences and commonalities in language use per role. The commonalities refer to tasks, e.g. managing a project, additional roles, such as being a project leader, and genre, i.e. writing for the purpose of acquiring research funding. The roles differ in the research areas that people are involved with and the prevalence of shared topics among roles.

Several limitations apply: first, the application of the methodology presented herein to a public, real-world dataset is of illustrative nature. Thorough empirical work is needed to derive at conclusive statements about the similarities and differences between roles, and the relationship between network roles and language choice in this domain of research and innovation.

Second, our approach is very coarse in that it groups individuals who work in multiple different disciplines into a total of two different roles based on their structural characteristics. Yet, some meaningful differences between change agents and preservation agents were suggested by the data on this high level of aggregation. For future work, we plan to group people from similar domains into aggregates per role in order to test for differences on a more fine-grained level.

TABLE V. Most likely topics per role, key terms per topic, and self-defined topic labels

	change agents											
topic	project management	networking and	project management	regional developmen	waste management	engineering	energy	pollution	emission	public health	regional development	medical
1st	project	research	data	regional	water	structures	energy	water	engine	food	services	tnf
2nd	development	european	management	policy	waste	aircraft	gas	monitoring	diesel	europe	ict	disease
3rd	systems	europe	assessment	regions	european	material	hydrogen	eu	combustion	human	business	gene
4th	system	network	tools	policies	europe	materials	combustion	chemical	fuel	virus	satellite	arthritis
5th	based	innovation	project	development	land	performance	biomass	pollutants	sensor	studies	rural	human
6th	high	knowledge	information	sustainable	market	composite	solar	directive	emission	million	information	mouse
7th	develop	training	fisheries	region	eu	damping	fuel	system	integrated	developing	robot	genes
8th	technologies	projects	support	national	smes	forming	low	pollution	power	health	communication	diseases
9th	control	support	studies	sustainability	aquaculture	monitoring	process	groundwater	emissions	forest	systems	mice
DP	0.731	0.276	0.165	0.080	0.070	0.055	0.053	0.050	0.046	0.044	0.038	0.036
	preservation agents											
topic	project management	research in EU	industry	networking and learning	environment	genetics	energy	transportati on	cancer	security	industry	public health
1st	project	research	production	research	water	genetic	energy	services	drug	governance	materials	food
2nd	european	european	products	network	management	gene	environmental	transport	clinical	security	properties	consumer
3rd	development	activities	industry	european	risk	genes	eu	solutions	cancer	social	devices	quality
4th	develop	countries	design	excellence	environmenta	disease	policy	business	cell	science	temperature	products
5th	research	information	manufacturing	integration	data	genomic	assessment	information	cells	eu	techniques	production
6th	systems	eu	product	training	monitoring	factors	agricultural	cities	hiv	issues	high	animal
7th	based	projects	industrial	europe	information	molecular	european	end	tumour	public	industrial	safety
8th	integrated	europe	processes	knowledge	assessment	genomics	sustainable	service	therapeutic	ethical	based	health
9th	knowledge	action	materials	researchers	practices	studies	impact	data	molecular	europe	structures	project
DP	0.921	0.414	0.160	0.102	0.080	0.077	0.076	0.071	0.062	0.061	0.056	0.055

Third, the dataset we used might be incomplete. Based on public information we cannot assess how many projects are not listed. Furthermore, the CORDIS database does not include rejected proposals. Also, the techniques we used for merging instances of names leaves room for improvement: errors such as typos could be further eliminated by employing edit-distance algorithms. Variations in names due to name changes, such as in the case of women adopting their husband's last name, might require careful processing of the institutional affiliation and address fields.

Forth, the operationalization of roles via certain network analytical measures and respective ranges of values for is debatable. A smaller set of metrics might lead to the different results, and so might the inclusion of additional metrics. We plan to investigate the robustness of our findings depending on various metrics and values in the future. Furthermore, we will include temporal and geospatial information for identifying strong and weak ties. These types of meta-data are available in our dataset. Furthermore, we plan to investigate the relationship between network roles, language use, and the amount of funding awarded.

Fifth, the process of identifying individuals per role is not fully automated and requires basic expertise in network analysis. Even though we are working on further automating this process, such expertise can be useful in interpreting the topic modeling results.

Finally, the network roles of change agents and preservation agents are not necessarily exclusive. For example, a person could direct a department or a group at her institution where she establishes a certain research agenda that is persistently followed (preservation agent). At the same time, the same person can be a boundary spanner connecting various external groups and their ideas, e.g. by serving on program committees and review boards (change agent). For this project, we retrieved all text files for each person identified per role. Since these roles are not exclusive, the retrieved texts per role might also overlap, and they did for this dataset. Consequently, the text data are exclusive across FPs, but not within FPs. In future work, we will separate the texts per person depending on what role that person played per project. This enhancement of the presented methodology allows for accounting for the fact that individuals can play different roles in different networks, on different teams, and for different tasks. In future work, we will furthermore focus on measuring the diffusion of topics addressed by either role across time, i.e. subsequent Framework Programs, and across roles.

ACKNOWLEDGMENT

We are grateful to Dr. Carolyn Penstein Rosé from Carnegie Mellon University for her advice on this project.

REFERENCES

[1] D. Alderson, "Catching the'network science'bug: Insight and opportunity for the operations researcher," *Operations Research,* vol. 56, pp. 1047-1065, 2008.

[2] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on Enron and academic email," *Journal of Artificial Intelligence Research,* vol. 30, pp. 249-272, 2007.

[3] E. Adar and L. Adamic, "Tracking Information Epidemics in Blogspace," in *2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne, France, 2005, pp. 207-214.

[4] C. Roth and J. Cointet, "Social and semantic coevolution in knowledge networks," *Social Networks,* 2009.

[5] S. R. Corman, T. Kuhn, R. D. McPhee, and K. J. Dooley, "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication," *Human Communication Research,* vol. 28, pp. 157-206, 2002.

[6] J. A. Danowski, "Network Analysis of Message Content," *Progress in Communication Sciences,* vol. 12, pp. 198-221, 1993.

[7] J. Milroy and L. Milroy, "Linguistic change, social network and speaker innovation," *Journal of Linguistics,* vol. 21, pp. 339-384, 1985.

[8] T. Griffiths, M. Steyvers, and J. Tenenbaum, "Topics in semantic representation," *Psychological Review,* vol. 114, pp. 211-244, 2007.

[9] M. Granovetter, "The strength of weak ties," *American journal of sociology,* vol. 78, pp. 1360-1380, 1973.

[10] E. Rogers, *Diffusion of Innovations*. Glencoe: Free Press, 1962.

[11] D. J. Watts, "The Accidental Influentials," *Harvard Business Review,* pp. 22-23, 2007.

[12] I. McAllister and D. Studlar, "Bandwagon, underdog, or projection?: Opinion polls and electoral choice in Britain, 1979-1987," *Journal of Politics,* vol. 16, 1991.

[13] J. Chang, J. Boyd-Graber, and D. Blei, "Connections between the lines: augmenting social networks with text," in *15th ACM SIGKDD International Conference*, Paris, France, 2009.

[14] D. Mimno and A. McCallum, "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression," in *24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, Helsinki, Finland, 2008.

[15] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[16] CORDIS, "Community Research and Development Information Service ".

[17] J. Diesner and K. M. Carley, "He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. ," in *IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA)*, Ottawa, Canada, 2009.

[18] K. M. Carley, J. Reminga, J. Storrick, and M. DeReno, "ORA User's Guide 2009," Carnegie Mellon University, School of Computer Science, Institute for Software Research Technical Report CMU-ISR-09-115, 2009.

[19] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002.

[20] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2009.