# Social Network Extraction from Texts: A Thesis Proposal

**Apoorv Agarwal**
Department of Computer Science
Columbia University
apoorv@cs.columbia.edu

## Abstract

In my thesis, I propose to build a system that would enable extraction of social interactions from texts. To date I have defined a comprehensive set of social events and built a preliminary system that extracts social events from news articles. I plan to improve the performance of my current system by incorporating semantic information. Using domain adaptation techniques, I propose to apply my system to a wide range of genres. By extracting linguistic constructs relevant to social interactions, I will be able to empirically analyze different kinds of linguistic constructs that people use to express social interactions. Lastly, I will attempt to make convolution kernels more scalable and interpretable.

## 1 Introduction

Language is the primary tool that people use for establishing, maintaining and expressing social relations. This makes language the real carrier of social networks. The overall goal of my thesis is to build a system that automatically extracts a social network from raw texts such as literary texts, emails, blog comments and news articles. I take a "social network" to be a network consisting of individual human beings and groups of human beings who are connected to each other through various relationships by the virtue of participating in *social events*. I define social events to be events that occur between people where at least one person is aware of the other and of the event taking place. For example, in the sentence *John talks to Mary*, entities *John* and *Mary* are aware of each other and of the talking event. In the sentence *John thinks Mary is great*, only *John* is aware of *Mary* and the event is the thinking event. My thesis will introduce a novel way of constructing networks by analyzing text to capture such interactions or events.

Motivation: Typically researchers construct a social network from various forms of electronic interaction records like self-declared friendship links, sender-receiver email links and phone logs etc. They ignore a vastly rich network present in the content of such sources. Secondly, many rich sources of social networks remain untouched simply because there is no meta-data associated with them (literary texts, new stories, historical texts). By providing a methodology for analyzing language to extract interaction links between people, my work will overcome both these limitations. Moreover, by empirically analyzing large corpora of text from different genres, my work will aid in formulating a comprehensive linguistic theory about the types of linguistic constructs people often use to interact and express their social interactions with others. In the following paragraphs I will explicate these impacts.

**Impact on current SNA applications**: Some of the current social network analysis (SNA) applications that utilize interaction meta-data to construct the underlying social network are discussed by Domingos and Richardson (2003), Kempe et al. (2003), He et al. (2006), Rowe et al. (2007), Lindamood et al. (2009), Zheleva and Getoor (2009). But meta-data captures only part of all the interactions in which people participate. There is a vastly rich network present in text such as the content of emails, comment threads on online social networks, transcribed phone calls. My work will enrich the

social network that SNA community currently uses by complementing it with the finer interaction linkages present in text. For example, Rowe et al. (2007) use the sender-receiver email links to connect people in the Enron email corpus. Using this network, they predict the organizational hierarchy of the Enron Corporation. Their social network analysis for calculating centrality measure of people does not take into account interactions that people talk about in the content of emails. Such linkages are relevant to the task for two reasons. First, people talk about their interactions with other people in the content of emails. By ignoring these interaction linkages, the underlying communication network used by Rowe et al. (2007) to calculate various features is incomplete. Second, sender-receiver email links only represent "who talks to whom". They do not represent "who talks about whom to whom." This later information seems to be crucial to the task presumably because people at the lower organizational hierarchy are more likely to talk about people higher in the hierarchy. My work will enable extraction of these missing linkages and hence offers the potential to improve the performance of currently used SNA algorithms. By capturing alternate forms of communications, my system will also overcome a known limitation of the Enron email corpus that a significant number of emails were lost at the time of data creation (Carenini et al., 2005).

**Impact on study of literary and journalistic texts**: Sources of social networks that are primarily textual in nature such as literary texts, historical texts, or news articles are currently under-utilized for social network analysis. In fact, to the best of my knowledge, there is no formal comprehensive categorization of social interactions. An early effort to illustrate the importance of such linkages is by Moretti (2005). In his book, *Graphs, Maps, Trees: Abstract Models for a Literary History*, Moretti presents interesting insights into a novel by looking at its interaction graph. He notes that his models are incomplete because they neither have a notion of weight (number of times two characters interact) nor a notion of direction (mutual or one-directional). There has been recent work that partially addresses these concerns (Elson et al., 2010; Celikyilmaz et al., 2010). They only extract mutual interactions that are signaled by quoted speech. My thesis will go beyond quoted speech and will extract interactions signaled by any linguistic means, in particular verbs of social interaction. Moreover, my research will not only enable extraction of mutual linkages ("who talks to whom" ) but also of one-directional linkages ("who talks about whom"). This will give rise to new applications such as characterization of literary texts based on the type of social network that underlies the narrative. Moreover, analyses of large amounts of related text such as decades of news articles or historical texts will become possible. By looking at the overall social structure the analyst or scientist will get a summary of the key players and their interactions with each other and the rest of network.

**Impact on Linguistics**: To the best of my knowledge, there is no cognitive or linguistic theory that explains how people use language to express social interactions. A system that detects lexical items and syntactic constructions that realize interactions and then classifies them into one of the categories, I define in Section 2, has the potential to provide linguists with empirical data to formulate such a theory. For example, the notion of social interactions could be added to the FrameNet resource (Baker and Fillmore, 1998) which is based on frame semantics. FrameNet records possible semantic frames for lexical items. Frames describe lexical meaning by specifying a set of frame elements, which are participants in a typical event or state of affairs expressed by the frame. It provides lexicographic example annotations that illustrate how frames and frame elements can be realized by syntactic constructions. My categorization of social events can be incorporated into FrameNet by adding new frames for social events to the frame hierarchy. The data I collect using the system can provide example sentenctes for these frames. Linguists can use this data to make generalizations about linguistic constructions that realize social interactions frames. For example, a possible generalization could be that transitive verbs in which both subject and object are people, frequently express a social event. In addition, it would be interesting to see what kind social interactions occur in different text genres and if they are realized differently. For example, in a news corpus we hardly found expressions of non-verbal mutual interactions (like eye-contact) while these are frequent in fiction

texts like *Alice in Wonderland*.

## 2 Work to date

So far, I have defined a comprehensive set of social events and have acquired reliable annotations on a well-known news corpus. I have built a preliminary system that extracts social events from news articles. I will now expand on each of these in the following paragraphs.

**Meaning of social events**: A text can describe a social network in two ways: explicitly, by stating the type of relationship between two individuals (e.g. *Mary is John's wife*), or implicitly, by describing an event which initiates or perpetuates a social relationship (e.g. *John talked to Mary*). I call the later types of events "social events" (Agarwal et al., 2010). I defined two broad types of social events: **interaction**, in which both parties are aware of each other and of the social event, e.g., a conversation, and **observation**, in which only one party is aware of the other and of the interaction, e.g., thinking of or talking about someone. For example, sentence 1, contains two distinct social events: interaction: *Toujan* was informed by the *committee*, and observation: *Toujan* is talking about the *committee*. I have also defined sub-categories for each of these broad categories based on physical proximity, verbal and non-verbal interactions. For details and examples of these sub-categories please refer to Agarwal et al. (2010)

(1) [Toujan Faisal], 54, {said} [she] was {informed} of the refusal by an [Interior Ministry committee] overseeing election preparations.

As a pilot test to see if creating a social network based on social events can give insight into the social structures of a story, I manually annotated a short version of *Alice in Wonderland*. On the manually extracted network, I ran social network analysis algorithms to answer questions like: who are the most influential characters in the story, which characters have the same social roles and positions. The most influential characters in the story were detected correctly. Another finding was that characters appearing in the same scene like *Dodo*, *Lory*, *Eaglet*, *Mouse* and *Duck* were assigned the same social roles and positions. This pointed out the possibility of using my method to identify separate scenes or sub-plots in a narrative, which is crucial for a better understanding of the text under investigation.

Motivated by this pilot test I decided to annotate social events on the Automatic Content Extraction (ACE) dataset (Doddington et al., 2004), a well known news corpus. My annotations extend previous annotations for entities, relations and events that are present in the 2005 version of the corpus. My annotations revealed that about 80% of the times, entities mentioned together in the same sentence were not linked with any social event. Therefore, a simple heuristic of connecting entities that are present in the same sentence with a link will not reveal a meaningful network. Hence I saw a need for a more sophisticated analysis.

**Extraction of social events**: To perform such an analysis, I built models for two tasks: social event detection and social event classification (Agarwal and Rambow, 2010). Both were formulated as binary tasks: the first one being about detecting existence of a social event between a pair of entities in a sentence and the second one being about differentiating between the interaction and observation type events (given there is an event between the entities). I used tree kernels on structures derived from phrase structure trees and dependency trees in conjunction with Support Vector Machines (SVMs) to solve the tasks. For the design of structures and type of kernel, I took motivation from a system proposed by Nguyen et al. (2009) which is a state-of-the-art system for relation extraction. I tried all the kernels and their combinations proposed by Nguyen et al. (2009). I used syntactic and semantic insights to devise a new structure derived from dependency trees and showed that this plays a role in achieving the best performance for both social event detection and classification tasks. The reason for choosing such representations is motivated by extensive studies about the regular relation between verb alternations and meaning components (Levin, 1993; Schuler, 2005). This regularity provides a useful generalization that helps to overcome lexical sparseness. However, in order to exploit such regularities, there is a need to have access to a representation which makes the predicate-argument structure clear. Dependency representations do this. Phrase structure representations also represent predicate-argument structure,

but in an indirect way through the structural configurations. These experiments showed that as a result of how language expresses the relevant information, dependency-based structures are best suited for encoding this information. Furthermore, because of the complexity of the task, a combination of phrase-based structures and dependency-based structures perform the best. To my surprise, the system performed extremely well on a seemingly hard task of differentiating between interaction and observation type social events. This result showed that there are significant clues in the lexical and syntactic structures that help in differentiating mutual and one-directional interactions.

## 3 Future Work

Currently I am working on incorporating semantic resources to improve the performance of my preliminary system. I will work on making convolution kernels scalable and interpretable. These two steps will meet my goal of building a system that will extract social networks from news articles. My next step will be to survey and incorporate domain adaptation techniques that will allow me port my system to other genres like literary and historical texts, blog comments, emails etc. These steps will allow me to extract social networks from a wide range of textual data. At the same time I will be able to empirically analyze the types of linguistic patterns, both lexical and syntactic, that perpetuate social interactions. Now I will expand on the aforementioned future directions.

**Adding semantic information**: Currently I am exploring linguistically motivated enhancements of dependency and phrase structure trees to formulate new kernels. Specifically, I am exploring ways of incorporating semantic information from VerbNet and FrameNet. This will help me reduce data sparseness and thus improve my current system. I am interested in modeling classes of events which are characterized by the cognitive states of participants– who is aware of whom. The predicate-argument structure of verbs can encode much of this information very efficiently, and classes of verbs express their predicate-argument structure in similar ways. Levin's verb classes, and Palmer's VerbNet (Levin, 1993; Schuler, 2005), are based on syntactic similarity between verbs: two verbs are in the same class if and only if they can realize their arguments in the same syntactic patterns. By the Levin Hypothesis, this is because they share meaning elements, and meaning and syntactic realizations of arguments are related. However, this does not mean that verbs in the same Levin or VerbNet class are synonyms; for example, *to deliberate* and *to play* are both in VerbNet class meet-36.3-1. But from a social event perspective, I am not interested in exact synonymy, and in fact it is quite possible that what I am interested in (awareness of the interaction by the event participants) is the same among verbs of the same VerbNet class. In this case, VerbNet will provide a useful abstraction. Future work will also explore FrameNet, which provides a different type of semantic abstraction and explicit semantic relations that are not directly based on syntactic realizations.

**Scaling convolution kernels**: Convolution kernels, first proposed by Haussler (1999), are a convenient way of "naturally" combining a variety of features without having to do fine-grained feature engineering. Collins and Duffy (2002) presented a way of successfully using them for NLP tasks such as parsing and tagging. Since then they have been used for various NLP tasks such as relation extraction (Zelenko et al., 2002; Culotta and Jeffrey, 2004; Nguyen et al., 2009), semantic role labeling (Moschitti et al., 2008), question-answer classification (Moschitti et al., 2007) etc. Convolution kernels calculate the similarity between two objects, like trees or strings, by a recursive calculation over the "parts" (substrings, subtrees) of objects. This calculation is usually made computationally efficient by using dynamic programming. But there are two limitations: 1) the computation is still quadratic and hence slow and 2) the features (or parts) that are given high weights at the time of learning remain inaccessible i.e. interpretability of the model becomes difficult.

One direction I will explore to make convolution kernels more scalable is the following: The decision function for the classifier (SVM in dual form) is given in equation 1 (Burges, 1998, Eq 61). In this equation, $y_i$ denotes the class of the $i^{th}$ support vector ($s_i$), $\alpha_i$ denotes the Lagrange multiplier of $s_i$, $K(s_i, x)$ denotes the kernel similarity between $s_i$ and a test example $x$, $b$ denotes the bias. The kernel definition proposed by Collins and Duffy (2002) is given in equation 2, where $h_s(T)$ is the number of

times the $s^{th}$ subtree appears in tree $T$. The kernel function $K(T_1, T_2)$ therefore calculates the similarity between trees $T_1$ and $T_2$ by counting the common subtrees in them. By combining equations 1 and 2 I get equation 3 which can be re-written as equation 4.

$$f(x) = \sum_{i=1}^{N_s} \alpha_i y_i K(s_i, x) + b \quad (1)$$

$$K(T_1, T_2) = \sum_{s} h_s(T_1) h_s(T_2) \quad (2)$$

$$f(x) = \sum_{i=1}^{N_s} \alpha_i y_i \sum_{s} h_s(s_i) h_s(x) \quad (3)$$

$$f(x) = \sum_{s} \sum_{i=1}^{N_s} \alpha_i y_i h_s(s_i) h_s(x) \quad (4)$$

The motivation for exchanging these summation signs is that the contribution of larger subtrees to the kernel similarity is strictly less than the contribution of the smaller subtrees. I will investigate the possibility of approximating the decision function of SVM without having to compare all subtrees, in particular large subtrees. I will also investigate if this summation can be calculated in parallel to make the calculation more scalable. Pelossof and Ying (2010) have done recent work on speeding up the Perceptron by stopping the evaluation of features at an early stage if they have high confidence that the example will be classified correctly. Another relevant work to improve the scalability of linear classifiers is due to Clarkson et al. (2010). However, to the best of my knowledge, there is no work that addresses approximation of kernel evaluation for convolution kernels.

**Interpretability of convolution kernels**: As mentioned in the previous paragraph, another disadvantage of using convolution kernels is that interpretability of a model is difficult. Recently, Pighin and Moschitti (2009) proposed an algorithm to linearize convolution kernels. They show that by efficiently encoding the "relevant" fragments generated by tree kernels, it is possible to get insight into the substructures that were given high weights at the time of learning a model. But their system currently returns thousands of such fragments. I will investigate if there is a way of summarizing these fragments into a meaningful set of syntactic and lexical classes. By doing so I will be able to empirically see what types of linguistic constructs are used by people to express different types of social interactions thus aiding in formulating a theory of how people express social interactions.

**Domain adaptation**: To be able to extract social networks from literary and historical texts, I will explore domain adaptation techniques. A notable work in this direction is by Daumé III (2007). This work is especially useful for me because Daumé III presents a straightforward kernelized version of his domain adaptation approach which readily fits the machine learning paradigm I am using for my problem. I will explore the literature to see if better domain adaptation techniques have been suggested since then. Domain adaptation will conclude my overall goal of creating a system that can extract social networks from a wide variety of texts. I will then attempt to extract social networks from the increasing amount of text that is becoming machine readable.

**Sentiment Analysis**:[1] A natural step to try once I have linkages associated with snippets of text is sentiment analysis. I will use my previous work (Agarwal et al., 2009) on contextual phrase-level sentiment analysis to analyze snippets of text and add polarity to social event linkages. Sentiment analysis will make the social network representation even richer by indicating if people are connected with positive, negative or neutral sentiments. This will not only give us information about the protagonists and antagonists in the text but will also affect the analysis of flow of information through the network.

## Acknowledgments

This work was funded by NSF grant IIS-0713548. I would like to thank Dr. Owen Rambow and Daniel Bauer for useful discussions and feedback.

## References

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using

[1] I do not mention sentiment analysis anywhere else in my proposal since I will simply use my earlier work.

lexical affect scoring and syntactic n-grams. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32.

Apoorv Agarwal, Owen C. Rambow, and Rebecca J. Passonneau. 2010. Annotation scheme for social network extraction from text. In *Proceedings of the Fourth Linguistic Annotation Workshop*.

C. Baker and C. Fillmore. 1998. The berkeley framenet project. *Proceedings of the 17th international conference on Computational linguistics*, 1.

Chris Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*.

G. Carenini, R. T. Ng, and X. Zhou. 2005. Scalable discovery of hidden emails from large folders. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 544–549.

Asli Celikyilmaz, Dilek Hakkani-Tur, Hua He, Greg Kondrak, and Denilson Barbosa. 2010. The actor-topic model for extracting social networks in literary narrative. *NIPS Workshop: Machine Learning for Social Computing*.

K. L. Clarkson, E. Hazan, and D. P. Woodruff. 2010. Sublinear optimization for machine learning. *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 449 –457.

M. Collins and N. Duffy. 2002. Convolution kernels for natural language. In *Advances in neural information processing systems*.

Aron Culotta and Sorensen Jeffrey. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 423–429, Barcelona, Spain, July.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ace) program–tasks, data, and evaluation. *LREC*, pages 837–840.

P. Domingos and M. Richardson. 2003. Mining the network value of customers. *In Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, pages 57–66.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147.

David Haussler. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

Jianming He, Wesley W. Chu, and Zhenyu (Victor) Liu. 2006. Inferring privacy information from social networks. *Intelligence and Security Informatics*, pages 154–165.

Hal Daume III. 2007. Frustratingly easy domain adaptation. *Annual Meeting-Association For Computational Linguistics*.

D. Kempe, J. Kleinberg, and E. Tardos. 2003. Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation.* The University of Chicago Press.

J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. 2009. Inferring private information using social network dataset. *WWW*.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

A. Moschitti, S. Quarteroni, and R. Basili. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. *Proceedings of the 45th Conference of the Association for Computational Linguistics (ACL)*.

A. Moschitti, D. Pighin, and R. Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34.

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. *Conference on Empirical Methods in Natural Language Processing*.

Raphael Pelossof and Zhiliang Ying. 2010. The attentive perceptron. *CoRR*, abs/1009.5972.

D. Pighin and A. Moschitti. 2009. Reverse engineering of tree kernel feature spaces. *Proceedings of the Conference on EMNLP*, pages 111–120.

Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore J Stolfo. 2007. Automated social hierarchy detection through email network analysis. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 109–117.

Karin Kipper Schuler. 2005. *Verbnet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI3179808.

D. Zelenko, C. Aone, and A. Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the EMNLP*.

Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. *Proceedings of the 18th international conference on World wide web*, pages 531–540.