Research Statement of Tanay Kumar Saha

1 Introduction

My research interest lies in the emerging area of Natural Language Processing, Data Mining and Machine Learning which includes deep learning, data analysis as well as information retrieval.

In the era of Big Data, we have a huge amount of data in the form of structured (networks) and unstructured content (text). Mining this data for retrieving useful information for the benefit of the human race requires an understanding of the content or information that the data holds and the relationship among the different modes of data. To this end, my primary research goal is threefold: (1) understanding the semantics of both the natural language and short-form text (e.g., twitter, forum, product reviews) by learning to represent them in space or through devising *highly interpretable and unbiased models* methods for compositional semantics; (2) understanding the semantics of nodes and edges, i.e., how they are related to each other in higher-order of various kinds networks (e.g., social network, biological network, the network of words, the network of textual units); and (3) finally combine the knowledge learned from both modes to understand complex semantics and reason beyond that is explicitly stated in each form of data by using the semantic information in various downstream tasks, such as classification, clustering, summarization, and question answering and domains, such as Natural Language Processing (NLP), Natural Language Understanding (NLU), Precision Medicine and messages (logs) generated from Mechanical Systems i.e. Heterogeneous Log Analysis.

A significant part of my PhD research focuses on learning to understand the semantics of natural language text (e.g., words, sentences, paragraph) and the nodes and edges in the network. The number of applications targeting to use the semantic knowledge learned from these domains is growing. However, tools and techniques developed so far are not sufficient to understand complex semantics and higher-order reasoning. More sophisticated methods which can capture essential aspects of both forms of the data and combine the knowledge from each type of the data are of enormous importance. Apart from the above works, I have also worked on mining information from a set of networks (graph Mining), finding significant subnetwork from a single large network (motif finding) and performing name disambiguation in anonymous author network. I show that the algorithm I developed for graph mining can be applied to find essential structures when applied to the interfacial region of a set of oligomeric proteins and the algorithm for motif finding can be used to classify apps from google app store as malignant or benign.

2 Research Accomplishment

2.1 Representation Learning of Textual and Network Units

In classical machine learning, hand-designed features are used for learning a mapping from the features. However, recently, there is a surge of research in representation learning which aims to learn abstract features given the input. For textual domain, representation learning focuses on learning important semantic information from words, phrases, sentences or any arbitrary unit of texts, such as paragraphs and documents. The learned representation can be used in various data mining tasks, such as classification, clustering, summarization and many others. For networks, representation learned for nodes and edges has been used for tasks, such as link prediction, collective classification, and many others. In my PhD thesis, I have crafted methods for learning representation for both the sentences and the network nodes/edges.

Recent studies on learning distributed representations for *words* have shown that semantic relations between words (e.g., synonymy, hypernymy, hyponymy) encoded in semantic lexicons like

WordNet or Framenet can improve the quality of word vectors that are trained solely on unlabeled data. My works [1, 2] are reminiscent of this line of research with a couple of crucial differences. Firstly, I am interested in the representation of sentences as opposed to words, for the former such resources are not readily available. Secondly, my main goal is to incorporate extra-sentential context in some form of inter-sentence relations as opposed to semantic relations between words. These differences posit many new research challenges: (*i*) how can we obtain extra-sentential context that can capture semantic relations between sentences? (*ii*) how can we effectively exploit the inter-sentence relations in our representation learning model? We solve these problems either by retrofitting or regularizing using the context network [1] or through a joint model [2] which predicts and regularize based on the context network of sentences. The joint model [2] is generic in terms of context and different modes of data the model can handle. The software is freely available from github: `https://github.com/tksaha/con-s2v/tree/jointlearning`.

Apart from text domain, understanding the dynamics of an evolving network is an important research problem with numerous applications in various fields, including social network analysis, information retrieval, recommendation systems, and bioinformatics. An essential task towards this understanding is to predict the likelihood of a future association between a pair of nodes, knowing the current state of the network. This task is commonly known as the *link prediction* problem. This problem has been studied extensively by many researchers from a diverse set of disciplines. In [3], I propose DYLINK2VEC[1], a novel learning method for obtaining a feature representation of node-pair instances, which is specifically suitable for the task of link prediction in a dynamic network.

2.2 Total Recall

Total Recall is an important task in supervised text mining. The objective of this task is to find everything about an entity or about a research question. For example, in Technology assisted review (TAR), the objective is to find all the documents relevant to a request for production in a legal matter, whereas, Systematic Review (SR) involves formulating a research question, searching in multiple biomedical databases, identifying relevant Randomized Control Trails (RCTs) based on abstracts and titles (abstract screening) and then based on full texts of a subset thereof, assessing their methodological qualities, extracting different data elements and synthesize them, and finally reporting the conclusions on a particular review question. The first step in this process is to browse a large collection of abstracts and label them as relevant or non-relevant for the given research question. However, abstract screening is a cumbersome process, and thus a sophisticated methods to solve the problem is of great importance. In [4], I propose two novel methods for batch-mode active learning for TAR using SVM. The novelty of the proposed methods is manifested in the way they choose the new batch of unlabeled instances for extending the prevailing training dataset. In [5], I study an existing abstract screening platform, Rayyan (`https://rayyan.qcri.org/`) and propose a 5-star rating algorithm to improve the user experience in abstract screening process.

2.3 Network Mining

Frequent subgraph mining (FSM) is an important research task in Network Mining Domain. It has application in various disciplines, including cheminformatics for solving QSAR (Quantitative Structure Activity Relationship) task, and in bioinformatics for finding structural motifs. The main objective of FSM is finding subgraph patterns that are frequent across a collection of graphs. Another important task is to find subgraphs which are candidate for motif in a given network. For this, we need to count each topology's frequency in the input network as well as in many randomized networks. Counting a topology's frequency in a single network is a challenging task as it requires solving subgraph isomorphism, a known $\mathcal{NP}$-complete problem. As the size of the motif grows, the number of candidate motifs increases exponentially, and the task becomes more challenging. To cope with the enormous computation cost of exhaustive counting of the frequency of candidate motifs, researchers consider various sampling based methods that obtain an approximation of relative frequency measure (which we call concentration) over all the candidates of a given size. In both the cases, I propose a sampling based solution which is efficient and works better than state-of-art

[1] DYLINK2VEC stands for **Link to Vec**tor in a **Dy**namic network. The proposed methodologies maps node-pairs (links) in a dynamic network to a vector representation.

methods. In [6], I propose a method for frequent subgraph mining, called FS^3, that is based on sampling of subgraphs of a fixed size [2]. In [7], I propose method on finding concentration of prospective motifs using a novel sampling based method. In subsequent works, I show that the algorithm (FS^3) I developed for graph mining can be applied to find essential structures when applied to the interfacial region of a set of oligomeric proteins [8] and the algorithm for motif finding [7] can be used to classify apps from google app store as malignant or benign [9].

2.4 Name Disambiguation

Entity disambiguation is not a new problem. In fact, named entity disambiguation has been a long standing problem in the field of bibliometrics and library science. The key reason for this is that many distinct authors share the same name, specifically considering the fact that the first name of an author is typically written in abbreviated form in the citation of many scientific journals. Thus, bibliographic servers that maintain such data may mistakenly aggregate the records from different persons into a single entity. In [10, 11], I design the task of solving name entity disambiguation using only graph topological information. We propose a simple solution that is robust and it takes only a few seconds to disambiguate a given node in real-life academic collaboration networks. The proposed method returns a real-valued score to rank the vertices of a network based on their likelihood for being an ambiguous node. So the score can be used as a pre-filter for identifying a small set of ambiguous references for subsequent analysis with a full set of features. Besides, the score can also be used independently as a feature for classification based solutions for entity disambiguation. This work is motivated by the growing need of data analysis without violating the privacy of the actors in a social network.

3 Future Research Goals

Designing Metric for evaluating context quality: The model that I propose for learning sentence representation [1, 2] is generic in terms of context it can handle. However, the noisy context may raise issues, for example, in discourse context we use previous and next sentence as context, however, they can be noisy i.e. they may deviate from the topic and may add noise in the representation. Thus, devising a method which can measure quality of context is of huge importance. I wish to develop such methods not only in monolingual setting but also in multilingual setting. This research should create curated structured information alike WordNet or Framenet about the different form of context usage styles, more specifically, what could be considered as an important context information in various languages.

Designing dataset for topical sentence representation methods: In our work [1, 2], we use an evaluation setting which is very different from the traditional evaluation setting, and our labels for the sentences are created automatically. In future work, I wish to curate the information by creating labels using human as an annotator, i.e., I wish to create a dataset where each sentence will have a topic label. This dataset will help to evaluate representation learning methods for sentences that aim to capture topic of a paragraph or a document.

Open Domain Question Answering: In information retrieval, open domain question answering is an important open task. An open-domain question answering system aims at returning an answer in response to the user's query. The returned answer is in the form of short texts rather than a list of relevant documents. Alexa, Cortana, and Google home are some of the examples of intelligent systems which communicate with the human in a more natural way use the existing solution of the open domain Question Answering (QA) which uses word similarity-based measures. More sophisticated methods which can encode information from a curated form of *context* given existing knowledge bases like Wikipedia, and *represent* and relate sentences of the question and the supporting facts are of enormous importance. I plan to devise such kind of method by leveraging the knowledge and experience I will learn from the previous two proposed ideas in addition to my work on learning

[2]The name FS^3 should be read as *F-S-Cube*, which is a compressed representation of the 4-gram composed of the bold letters in **F**ixed **S**ize **S**ubgraph **S**ampler.

sentence representation using context information.

Designing interpretable and unbiased representation learning model for textual and network units: Recently evaluation methods for embedding based models are evaluated either over some similarity tasks (word similarity tasks) or downstream tasks such as classification, clustering, summarization in text domain and link prediction in the network domain. In each case, existing models discard the following questions: (1) when can I trust the model, i.e., how can we make sure the model is not introducing various forms bias such as gender, race?; (2) how can we justify or interpret the distances learned between any two entities such as words or nodes; (3) when such model fails?, and (3) how do I correct an error in distance or bias?. Moreover, any kind of deep learning based model faces the following challenges as outline by DARPA [3]: (1) why did you do that? (2) why not something else? (3) when do you succeed? (4) when do you fail? (5) when can I trust you? (6) How do I correct an error?. Recently,

Total Recall: For abstract screening process, I conjecture that instead of showing the entire abstracts with highlighted keywords for producing interpretable results, it is more meaningful to show the key sentences from an abstract that resulted in the prediction of the abstract as relevant. Identifying such key sentences needs methods for text summarization. Key sentences from text summarization can also be used as features for text clustering or classification. Currently, latent representation of sentences showing promising results in both the supervised and unsupervised settings, so a good latent representation of sentences which can capture the topic can be of huge interest. I plan to extend my work described in which improves the summarization result for standard single document summarization tasks in both the TAR and Systematic review domain to develop new methods with improved performance and better interpretable results.

Enriching Biomedical KB: In biomedicine, KBs such as Pathway Interaction Database (NCI-PID) are crucial for understanding complex diseases such as cancer and for advancing precision medicine. While these knowledge bases are often carefully curated, they are far from complete. In non-static domains, new facts become true or are discovered at a fast pace, making the manual expansion of knowledge bases impractical. Extracting relations from a text corpus or inferring facts from the relationships among known entities are thus important approaches for populating existing knowledge bases. The KB completion task is to predict the existence of links (s, r, t) that are not seen in the training knowledge base.

Long Term Goal: I am also interested in multidisciplinary research that goes beyond representation learning and graph mining. I have recently embarked on joint research projects involving multiple research groups, where my collaborator and I are investigating multilingual (e.g. Bangla) language processing problems. Bengali is the seventh most spoken native language in the world by population. However, research effort for understanding the Bengali language is very scarce compared to English and European languages such as, German and French. This is due to the fact that labeled data and parallel text for learning multilingual representation is scarce and in some cases totally unavailable. Some of the efforts that I like to attend to semantically understand the language.

4 Conclusion

I am highly motivated and well prepared to pursue an academic research career. I have published in various top data mining and bioinformatics journal and conferences and presented at multiple conferences. I am also the first author of 90% of all my works. I have served as an external reviewer for top-notch data mining conferences and journal such as KDD, TKDD, and AAAI. Besides academic research, I am also very familiar with research work in industrial research labs; it is a huge plus, as data mining is thriving in the industry at a much faster pace. I include a full list of professional activities in my resume.

[3] https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)\%20IJCAI-16\%20DLAI\%20WS.pdf

Finally, I have a strong passion for my research area, which I desire to solve the most challenging research problems. Apart from solving the problem, I also plan to distribute publicly the code and data related to the research. All of my codes are available through my github account: `https://github.com/tksaha`. I am a firm advocate of reproducible research and plan to follow it through my career. Considering all of the above, I believe that I shall be very successful as a faculty member. If you are interested to know more about my research, please don't hesitate to contact me. I also welcome you to browse my website: `https://tksaha.github.io/`.

References

[1] **Tanay Kumar Saha**, Shafiq Joty, Naeemul Hassan, and Mohammad Al Hasan. Learning sentence representation with context. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management, CIKM*, 2017.

[2] **Tanay Kumar Saha**, Shafiq Joty, and Mohammad Al Hasan. Con-s2v: A joint learning framework for incorporating extra-sentential context into sen2vec. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, 2017.

[3] Mahmudur Rahman, **Tanay Kumar Saha**, Mohammad Al Hasan, Kevin S. Xu, and Chandan K. Reddy. Dylink2vec: Effective feature representation for link prediction in dynamic networks, 2017.

[4] **Tanay Kumar Saha**, Mohammad Al Hasan, Chandler Burgess, Md Ahsan Habib, and Jeff Johnson. Batch-mode active learning for technology-assisted review. In *IEEE International Conference on Big Data*, pages 1134–1143, 2015.

[5] **Tanay Kumar Saha**, Mourad Ouzzani, Hossam Hammady, Ahmed K. Elmagarmid, and Mohammad Al Hasan. Study of methods for abstract screening in a systematic review platform, 2017.

[6] **Tanay Kumar Saha** and Mohammad Al Hasan. Fsˆ3: A sampling based method for top-k frequent subgraph mining. *Statistical Analysis and Data Mining*, 8(4):245–261, 2015.

[7] **Tanay Kumar Saha** and Mohammad Al Hasan. Finding network motifs using mcmc sampling. In *Complex Networks VI*, pages 13–24. Springer International Publishing, 2015.

[8] **Tanay Kumar Saha**, Ataur Katebi, Wajdi Dhifli, and Mohammad Al Hasan. Discovery of functional motifs from the interface region of oligomeric proteins using frequent subgraph mining, 2017.

[9] Wei Peng, Tianchong Gao, Devkishen Sisodia, **Tanay Kumar Saha**, Feng Li, and Mohammad Al Hasan. Acts: Extracting android app topological signature through graphlet sampling (acceptance rate: 29%). In *IEEE Conference on Communications and Network Security*, 2016.

[10] **Tanay Kumar Saha**, Baichuan Zhang, and Mohammad Al Hasan. Name disambiguation from link data in a collaboration graph using temporal and topological features. *Social Network Analysis and Mining*, 5(1):1–14, 2015.

[11] **Tanay Kumar Saha**, Baichuan Zhang, and Mohammad Al Hasan. Name disambiguation from link data in a collaboration graph. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 81–84, 2014.