A Collective Approach to Graph Identification [1]

Galileo Namata and Lise Getoor
University of Maryland
College Park, MD USA
{namatag,getoor}@cs.umd.edu
http://www.cs.umd.edu/~getoor

There is a growing amount of observational data describing networks— examples include social networks, communication networks, and biological networks. As the amount of available data increases, so has our interest in analyzing these networks in order to uncover (1) general laws that govern their structure and evolution, and (2) patterns and predictive models to develop better policies and practices. However, a fundamental challenge in dealing with this newly available observational data describing networks is that the data is often of dubious quality—it is noisy and incomplete—and before any analysis method can be applied, the data must be cleaned, missing information inferred and mistakes corrected. Skipping this cleaning step can lead to flawed conclusions for things as simple as degree distribution and centrality measures; for more complex analytic queries, the results are even more likely to be inaccurate and misleading.

In this paper, we introduce the notion of *graph identification*, which explicitly models the inference of a "cleaned" output graph from a noisy input graph. We show how graph identification can be thought of as a series of probabilistic graph transformations. This is done via a combination of component models, in which the component models construct the output graph by merging nodes in the input graph (entity resolution), adding and deleting edges (link prediction), and labeling nodes (collective classification). We then present a simple, general approach to constructing local classifiers for predicting when to make these graph modifications, and combining the inferences into an overall graph identification framework. The problem is extremely challenging because there are dependencies among the transformation; ignoring the dependencies leads to sub-optimal results and modeling the dependencies correctly is also non-trivial.

Graph identification is closely related to work in *information extraction* [12]; information extraction, however, traditionally infers structured output from unstructured data (e.g., newspaper articles, emails), while graph identification is specifically focused on inferring structured data (i.e., the cleaned graph) from other structured data (i.e., the noisy graph, perhaps produced from an information extraction process). There is significant prior work exploring the components of graph identification individually; representatives include work on collective classification [7, 5, 6, 13], link prediction [4, 10, 8], and entity resolution [1, 2, 14]. More recently, there is work that looks at various ways these tasks are inter-dependent and can be modeled jointly [15, 11, 16, 9, 3]. To our knowledge, however, previous work has not formulated the complex structured prediction problem as interacting components which collectively infer the graph via a collection of probabilistic graph transformations.

In addition to defining the problem and describing a component solution approach, we present a complete system for graph identification. We show how the performance of graph identification is sensitive to the intra- and inter-dependencies among inferences. We evaluate on two real-world

[1] **Topic: Datamining**
Preference: Oral/Poster
Presenter: Lise Getoor

citation networks, with varying degrees of noise, and present a summary of our results showing (1) the overall utility of combining all of the components and (2) some of subtleties involved.

References

[1] Omar Benjelloun, Hector Garcia-Molina, Qi Su, and Jennifer Widom. Swoosh: A generic approach to entity resolution. Technical report, Stanford University, 2005.

[2] Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *KDD*, 1, 2007.

[3] Indrajit Bhattacharya, Shantanu Godbole, and Sachindra Joshi. Structured entity identification and document categorization: two tasks with one joint model. In *KDD*, 2008.

[4] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[5] Qing Lu and Lise Getoor. Link-based classification. In *ICML*, 2003.

[6] Luke McDowell, Kalyan Moy Gupta, and David W. Aha. Cautious inference in collective classification. In *AAAI*, pages 596–601, 2007.

[7] J. Neville and D. Jensen. Iterative classification in relational data. In *AAAI Workshop on Learning Statistical Models from Relational Data*, 2000.

[8] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations Newsletter*, 7(2):23–30, 2005.

[9] Hoifung Poon and Pedro Domingos. Joint unsupervised coreference resolution with markov logic. In *EMNLP*, 2008.

[10] Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. In *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003.

[11] Dan Roth and Wen-Tau Yih. A linear programming formulation for global inference in natural language tasks. In *CoNLL*, 2004.

[12] Sunita Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3), 2008.

[13] Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[14] Parag Singla and Pedro Domingos. Entity resolution with markov logic. *ICDM*, 2006.

[15] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *Adv. in Neural Info. Proc. Sys.*, December 2003.

[16] Michael L. Wick, Khashayar Rohanimanesh, Karl Schultz, and Andrew McCallum. A unified approach for schema matching, coreference and canonicalization. In *KDD*, 2008.