Literature Search through Mixed-Membership Community Discovery

Tina Eliassi-Rad and Keith Henderson

Lawrence Livermore National Laboratory
P.O. Box 808, L-560, Livermore, CA 94551, USA
{eliassi, keith}@llnl.gov

Abstract. We introduce a new approach to literature search that is based on finding mixed-membership communities on an augmented co-authorship graph (ACA) with a scalable generative model. An ACA graph contains two types of edges: (1) coauthorship links and (2) links between researchers with substantial expertise overlap. Our solution eliminates the biases introduced by either looking at citations of a paper or doing a Web search. A case study on PubMed shows the benefits of our approach.

Keywords: Literature search, mixed-membership, community discovery.

1 Introduction

Given a research topic (e.g. reconstruction of the 1918 influenza virus) and a couple of seminal papers on that topic (e.g. [1] and [2]), how do we find authors who are conducting similar research? Traditional solutions to this problem include looking at the citations in the seminal papers and/or conducting Web searches on keywords associated with the chosen topic. Both of these commonly used solutions have biases that limit their effectiveness. For example, looking only at the citations of a paper provides a partial view of the domain (namely, the ones provided by the authors). Doing a Web search on keywords neglects the wealth of information embedded in social networks (such as co-authorship graphs).

In this work, we propose a new approach to the literature search problem that is based on finding mixed-membership communities on an augmented co-authorship (ACA) graph. We construct an ACA graph by fusing the information from a bipartite expertise-by-author graph into a co-authorship graph, which produces a denser and more structured version of the original co-authorship graph.

For the mixed-membership community discovery algorithm, we utilize our Latent Dirichlet Allocation for Graphs (LDA-G) [3]. LDA-G is a scalable generative model that adapts the Latent Dirichlet Allocation (LDA) [4] topic-modeling algorithm for use in graphs rather than text corpora. A simple post-analysis of

LDA-G's communities provides a ranking of the most similar authors. In our experiments on PubMed[1] data, LDA-G produces better solutions than when it is applied to regular co-authorship graphs or bipartite expertise-by-author graphs. In addition to our qualitative results, we provide quantitative results based on link prediction performance of LDA-G's posterior estimate.

2 Mixed-Membership Community Discovery

We utilize our scalable generative LDA-G model [3] to find mixed-membership communities in large graphs. In this context, "mixed membership" means that nodes can belong to multiple communities with varying probabilities. Given a graph, LDA-G models each source node in the graph as a multinomial distribution over some set of communities Z. The cardinality of Z is unknown a priori and is learned via Bayesian inference from a Dirichlet prior. In LDA-G, each source node generates a series of communities from its multinomial; and each community is a multinomial distribution over target nodes. Any time a community is generated by a source node, that community generates a target node from its distribution. The distributions over source-node to community and community to target-node are learned using MCMC techniques (e.g., we use Gibbs sampling). To simplify inference, it is assumed that the behaviors of a node as a source-node and as a target-node are probabilistically independent. The generative model for LDA-G is as follows:

$$t_i \mid z_i, \varphi^{(z_i)} \sim Discrete(\varphi^{(z_i)}) \tag{1}$$

$$\varphi \sim Dirichlet(\beta) \tag{2}$$

$$z_i \mid \theta^{s_i} \sim Discrete(\theta^{s_i}) \tag{3}$$

$$\theta \sim Dirichlet(\alpha) \tag{4}$$

Equations 1 and 3 are the multinomial distributions from communities z to target-nodes t and from source-nodes s to communities z, respectively. Equation 2 and 4 are the prior distributions on target nodes with hyperparameter β and on communities with hyperparameter α, respectively.

Unlike most approaches to community discovery, LDA-G only requires present links (i.e., non-zero entries in the adjacency matrix). This property helps its runtime and space complexities. It has $O(NKM)$ runtime and $O(N(K+M))$ space complexity, where N is the number of nodes in the graph, K is the number of communities ($K << N$), and M is the average vertex degree in the graph ($M << N$).

We define link-prediction performance as a quantitative way of measuring the effectiveness of LDA-G in factoring a graph into communities. In particular,

[1] PubMed is a repository containing millions of citations from biomedical articles (http://www.pubmedcentral.nih.gov/).

we compute area under ROC curve on the task of predicting links from held-out test-sets based on the (posterior) probability of a link between two nodes s and t. Equation 5 defines this probability.

$$p(s \rightarrow t) = \sum_{z \in Z} p(z|s)p(t|z) \tag{5}$$

There are a few scalable generative models that find community structure in graphs [3, 5–10]; most of them extend LDA. The simplest adaptions are LDA-G and SSN-LDA [9]. There are also derivations that find communities in social networks with weighted links [8] or with categorical attributes on links [5]; find communities in textual attributes and relations [6, 10]; and find communities in dynamic (time-evolving) graphs [7].

3 Augmented Co-authorship (ACA) Graph

An ACA graph is a denser and more structured version of a co-authorship graph. We construct an ACA graph by fusing the information from a bipartite expertise-by-author graph into a standard co-authorship graph. We advocate a two-step approach for the fusion. First, we prune the expertise-by-author multigraph[2] by removing links that appear less than r times (i.e., links with weights $< r$). We pick the threshold r based on the distribution of weights on the expertise-by-author links. This step effectively removes "noisy" and "random" links from the expertise-by-author graph. Second, in the co-authorship graph, we add a link between any pair of authors that share an expertise in the pruned expertise-by-author graph. Hence, the ACA graph not only contains co-authorship links but also links indicating that two authors have substantial overlap in their expertise.

The intuition behind ACA graphs is that fusing data from different sources, especially introducing more structured data into less structured data, can be quite valuable during analysis. Figure 1 depicts the adjacency matrices for an expertise-by-author graph and a co-authorship graph extracted from PubMed and their associated ACA graph. Table 1 presents the basic statistics of these data graphs. The expertise nodes were extracted based on term frequency in PubMed abstracts. A link exists from an expertise node x to an author node y for every paper in which y is an author and x is a term appearing in the paper's abstract.

To generate the ACA graph, we need to select a threshold r to remove "noisy" and "random" links from the expertise-by-author graph. In other words, we want only the expertise-by-author relationships that are "significant" (because we are going to generate implicit co-authorship links between authors with significant expertise overlap). Figure 2 depicts the distribution of edge weights on our

[2] Each time an author publishes in a given expertise, a link is created in the bipartite expertise-by-author graph.

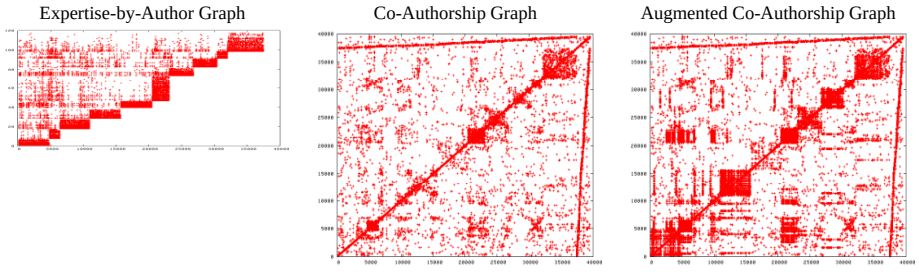

Fig. 1. Adjacency matrices for an expertise-by-author graph and a co-authorship graph extracted from PubMed and their augmented co-authorship graph. (The expertise-by-author graph's adjacency matrix is sorted by the order in which each author's expertise was added to the graph.)

Table 1. Basic statistics on our PubMed-extracted data graphs (LCC is short for the largest connected component).

Data Graph	# of Nodes	# of Links	# of Components	% of Nodes in LCC	% of Links in LCC
Expertise-by-Author	117 (E) 37,483 (A)	119,443	1	100%	100%
Co-Authorship	37,227 (A)	143,364	4,556	23.54%	35.82%
Augmented Co-Authorship	37,227 (A)	339,644	4,389	30.40%	46.89%

expertise-by-author graph. We used a threshold r of 12 for in our case-study. The probability of an edge weight being greater than or equal to 12 is 1.3%; hence, the links associated these weights do not exist because of chance or noise. Our threshold generated a pruned expertise-by-author graph with 1,310 authors (3.5% of the original authors), 117 expertise terms, and 1,565 links (1.3% of the original expertise-by-author links).

4 Experiments

Given the graphs depicted in Figure 1, we find mixed-membership communities on them with LDA-G, and then use the community structures to find authors that are performing similar research to authors of [1] and [2] (i.e. research on the reconstruction of the 1918 influenza virus). For the latter, we look for communities that are common between authors of [1] and [2]. In all three graphs, LDA-G finds communities that are common between authors of [1] and [2]. In the expertise-by-author graph, LDA-G finds four common communities (see Figure 3, top plot, communities #10, #20, #32, and #33). In the co-authorship graph, it uncovers one common community (see Figure 3, middle plot, community #6). In the ACA graph, it discovers three common communities (see Figure 3, bottom plot, community #3, #14, and #16). It is only in the ACA

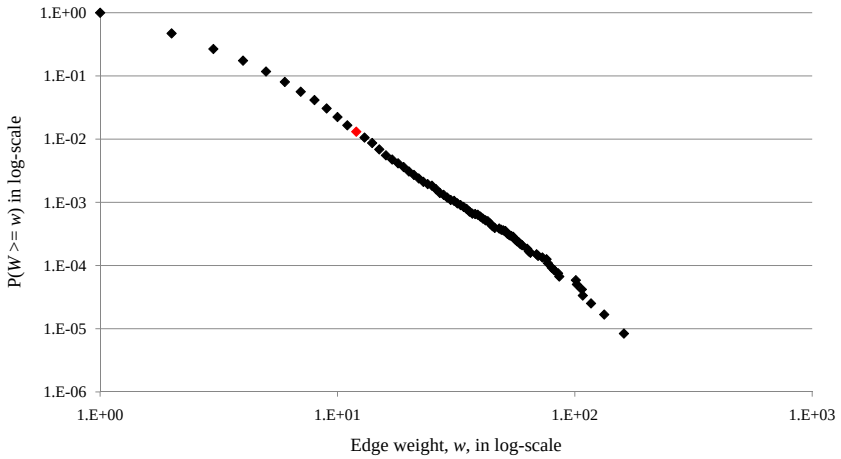

Fig. 2. Cumulative edge-weight distribution for the expertise-by-author graph. The chosen threshold ($r = 12$) is colored in red.

graph that LDA-G is able to find a common community with a significant overlap - specifically, 47% of authors of [2] and 30% of authors of [1] fall into community #3 of the ACA graph. Further inspection of this community reveals authors that have both similar co-authorship patterns and expertise as authors of [1] and [2]. We depict these authors and their expertise in the Figure 4. These authors have the highest percentage of membership in community #3 of the ACA graph, which is shared among authors of [1] and [2]. None of these authors were cited in [1] or [2]. We showed our findings to domain experts and received validation from them that we had indeed found the relevant researchers.

Figure 5 depicts the overlap in the expertise terms for authors of [1] and [2]. Even though both papers are on the reconstruction of 1918 influenza virus, the probability distribution on the expertise terms of their major author groups is different. In other words, simply conducting a keyword search on the (expertise) terms will not be sufficient for finding authors who are conducting similar research. LDA-G is able to effectively factor out a graph's community structure. Figure 6 plots the adjacency matrix and the resultant community-sorted matrix for the ACA graph. As it can be seen, LDA-G discovers nicely separated block-structure.

On link prediction, LDA-G's posterior estimates on the aforementioned graphs produce average area under the ROC curve (AUC) values of at least 0.918. (Recall that an AUC of 0.5 is a random guess.) Table 2 lists the AUC values on the PubMed graphs (averaged over 5 trials). As is standard in machine learning, we repeatedly divide the dataset into training and test sets, build a model on the training set, and examine its performance with respect to the chosen metric (e.g., AUC) on the held-out test-set. In particular, we use stratified random sampling to hold-out 1000 links from each graph. The remaining links are used to dis-

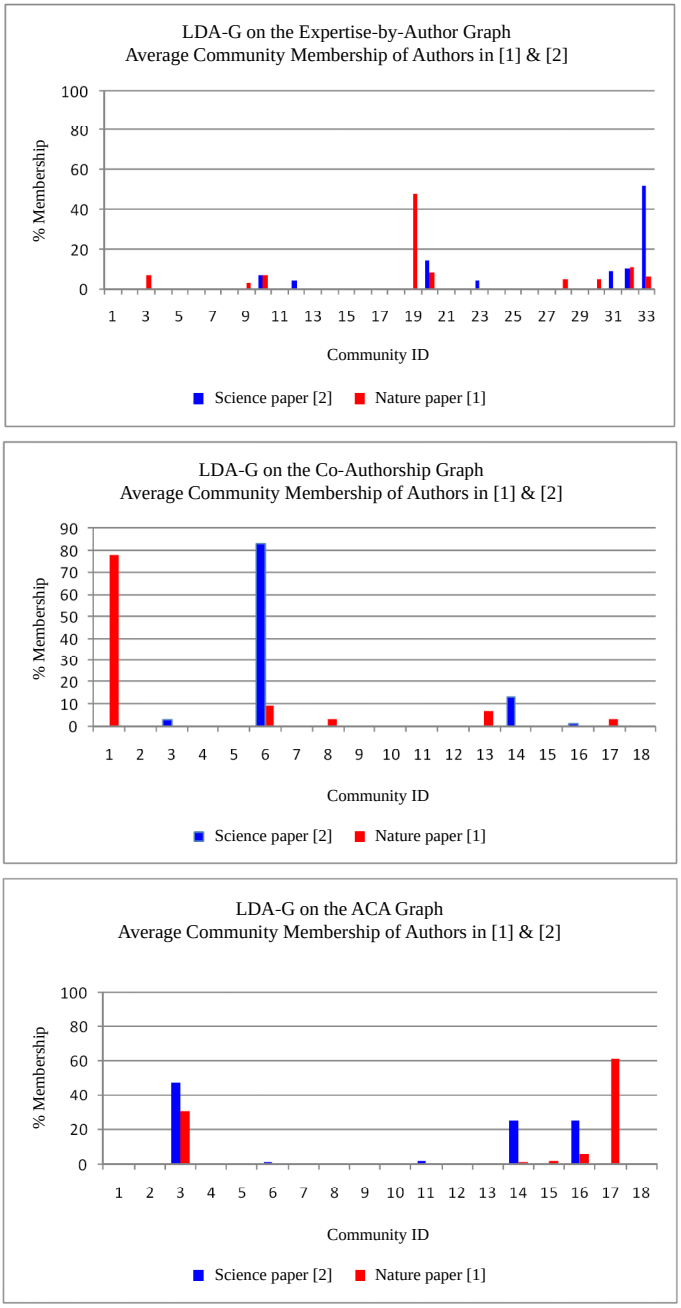

Fig. 3. Average community membership of authors of [1] and [2]. Only in the ACA graph do we find a common community (#3) with significant overlap between the authors of [1] and [2].

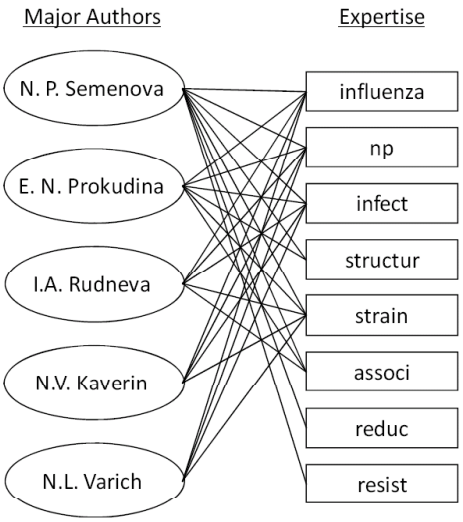

Fig. 4. Authors with the highest percentage of membership in community #3 of the ACA graph. These five authors and the authors of [1] and [2] share similar expertise and have topologically similar co-authorship neighborhoods.

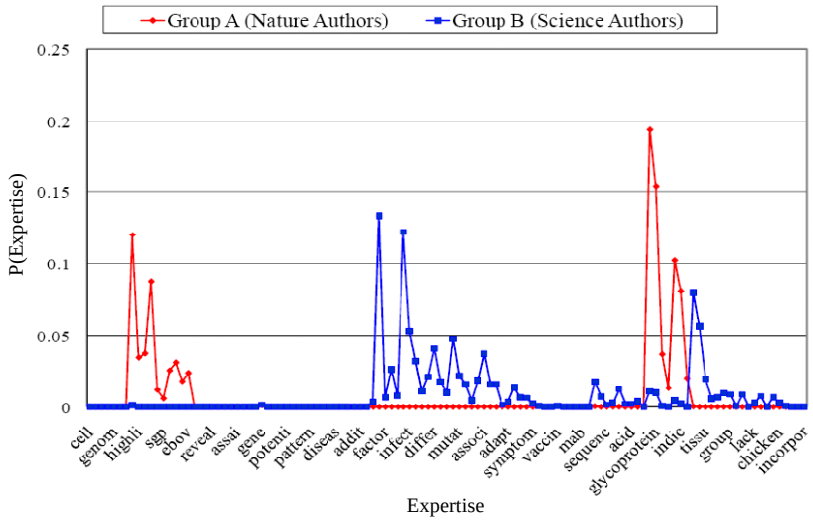

Fig. 5. LDA-G's qualitative results on the expertise-by-author graph. Plot shows the probability of expertise terms for major author groups of [1] in red and [2] in blue. Even though both papers are on reconstruction of the 1918 flu virus the authors' expertise terms does not overlap as much as expected.

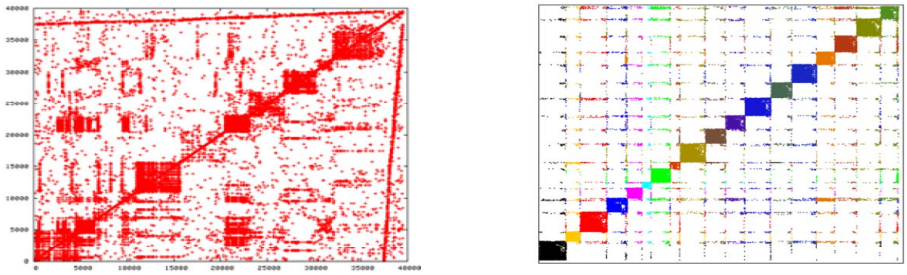

Fig. 6. The ACA Graph: Its adjacency matrix and its community-sorted matrix.

cover the latent communities. Then, the superiority of the discovery community structure is checked based on how well it predicts the existence of the held-out links as described in Equation 5. In [3, 5], we present a comparative study on link prediction results (on these graphs) between LDA-G and five other community discovery approaches (including *Fast Modularity*, *Cross Associations*, and *Infinite Relational Models*). LDA-G's link prediction results either outperform or are competitive with the best performer.

Table 2. AUC values on link prediction averaged over 5 trials (default value is 0.5).

Data Graph	LDA-G's Posterior Estimates
Expertise-by-Author	0.955
Co-Authorship	0.925
Augmented Co-Authorship	0.918

5 Conclusions

We describe a new approach to the literature search problem, which involves finding mixed membership communities on augmented co-authorship (ACA) graphs with LDA-G (a scalable generative model). An ACA graph contains not only co-authorship links but also links between researchers with substantial expertise overlap. We evaluate our approach qualitatively and quantitatively on data from PubMed and present a successful case study.

Future work involves utilizing the distributed-inference, temporal version of our LDA-G on larger-scale dynamic graphs in order to track the delineation of scientific domains/communities.

Acknowledgements. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48 and No. DE-AC52-07NA27344.

References

1. Kobasa, D., Jones, S.M., Shinya, K., Kash, J.C., Copps, J., Ebihara, H., Hatta, Y., Kim, J.H., Halfmann, P., Hatta, M., Feldmann, F., Alimonti, J.B., Fernando, L., Li, Y., Katze, M.G., Feldmann, H., Kawaoka, Y.: Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. Nature **445**(7125) (2007) 319–323
2. Tumpey, T.M., Basler, C.F., Aguilar, P.V., Zeng, H., Solórzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenberger, J.K., Palese, P., García-Sastre, A.: Characterization of the reconstructed 1918 spanish influenza pandemic virus. Science **310**(5745) (2005) 77–80
3. Henderson, K., Eliassi-Rad, T.: Applying latent Dirichlet allocation to group discovery in large graphs. In: Proceedings of the 24th Annual ACM Symposium on Applied Computing (SAC'09), Honolulu, HI (2009) 1456–1461
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
5. Henderson, K., Eliassi-Rad, T., Papadimitriou, S., Faloutsos, C.: Hcdf: A hybrid community discovery framework. In: Proceedings of the 2010 SIAM Conference on Data Mining (SDM'10), Columbus, OH (2010)
6. Li, H., Nie, Z., Lee, W.C., Giles, C.L., Wen, J.R.: Scalable community discovery on textual data with relations. In: Proceeding of the 17th ACM conference on Information and Knowledge Management (CIKM'08), Napa Valley, CA (2008) 1203–1212
7. Miller, K.T., Eliassi-Rad, T.: Continuous time group discovery in dynamic graphs. In: Notes of the 2009 NIPS Workshop on Analyzing Networks and Learning with Graphs, Whistler, BC, Canada (2009)
8. Zhang, H., Giles, C.L., Foley, H.C., Yen, J.: Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI'07), Vancouver, BC, Canada (2007) 663–668
9. Zhang, H., Qiu, B., Giles, C.L., Foley, H.C., Yen, J.: An LDA-based community structure discovery approach for large-scale social networks. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI'07), New Brunswick, NJ (2007) 200–207
10. Zhou, D., Manavoglu, E., Li, J., Giles, C.L., Zha, H.: Probabilistic models for discovering e-communities. In: Proceedings of the 15th international conference on World Wide Web (WWW'06), Edinburgh, Scotland (2006) 173–182