Literature Search through Mixed-Membership Community Discovery[*]

Tina Eliassi-Rad and Keith Henderson
Lawrence Livermore National Laboratory
{eliassirad1, henderson43}@llnl.gov

Introduction

Given a research topic (e.g. reconstruction of the 1918 influenza virus) and a couple of seminal papers on that topic (e.g. [4] and [6]), 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 our mixed-membership community discovery algorithm, we utilize our *Latent Dirichlet Allocation for Graphs* (*LDA-G*) [2]. LDA-G is a scalable generative model that adapts the Latent Dirichlet Allocation (LDA) [1] 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.

Mixed-Membership Community Discovery

We utilize our scalable generative LDA-G model [2] 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:

$$t_i | z_i, \varphi^{(z_i)} \sim Discrete(\varphi^{(z_i)}) \qquad \text{Multinomial from communities } z \text{ to target-nodes } t \qquad (1)$$

$$\varphi \sim Dirichlet(\beta) \qquad \text{Prior on target nodes with hyperparameter } \beta \qquad (2)$$

$$z_i | \theta^{s_i} \sim Discrete(\theta^{s_i}) \qquad \text{Multinomial from source-nodes } s \text{ to communities } z \qquad (3)$$

$$\theta \sim Dirichlet(\alpha) \qquad \text{Prior on communities with hyperparameter } \alpha \qquad (4)$$

---

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

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 ($K \ll N$), and M is the average vertex degree in the graph ($M \ll N$).

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

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

There are a few scalable generative models that find community structure in graphs [2, 3, 5, 7, 8, 9]; most of them extend LDA. The simplest adaptions are LDA-G and SSN-LDA [8]. There are also derivations that find communities in social networks with weighted links [7] or with categorical attributes on links [3]; or find communities in textual attributes and relations [5, 9].

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. We pick the threshold r based on graph-size considerations. This step effectively removes noise in 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[3] extracted from PubMed and their associated ACA graph. The expertise nodes here 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. We used a threshold r of 12 to remove noisy links from the expertise-by-author graph.

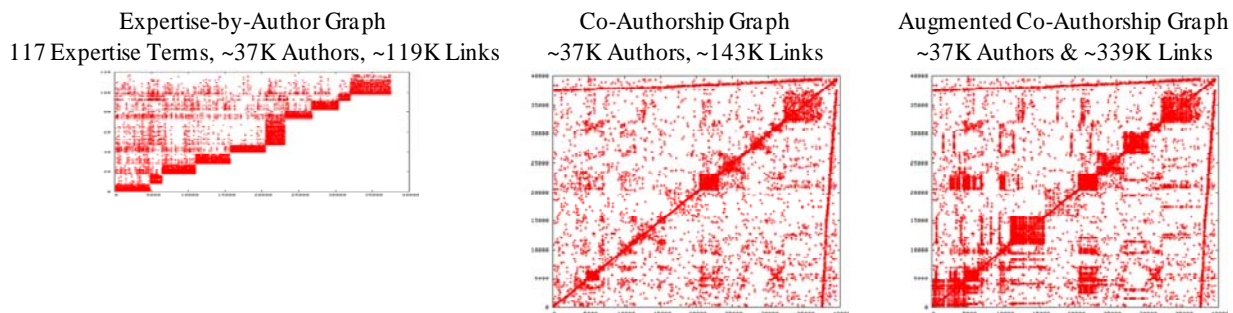| Expertise-by-Author Graph	Co-Authorship Graph	Augmented Co-Authorship Graph
117 Expertise Terms, ~37K Authors, ~119K Links	~37K Authors, ~143K Links	~37K Authors & ~339K Links

Figure 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.)

---

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

[3] This PubMed co-authorship graph is composed of 4555 connected components. The largest connected component has 8763 authors (approximately 24% of the entire graph).

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 [4] and [6] (i.e. research on the reconstruction of the 1918 influenza virus). For the latter, we look for communities that are common between authors of [4] and [6]. In all three graphs, LDA-G finds communities that are common between authors of [4] and [6]. In the expertise-by-author graph, LDA-G finds four common communities (see Figure 2, top left plot, communities #10, #20, #32, and #33). In the co-authorship graph, it uncovers one common community (see Figure 2, top right, community #6). In the ACA graph, it discovers three common communities (see Figure 2, bottom left plot, community #3, #14, and #16). It is only in the ACA graph that LDA-G is able to find a common community with a significant overlap – specifically, 47% of authors of [6] and 30% of authors of [4] 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 [4] and [6]. We depict these authors and their expertise in the Figure 2 (bottom right plot). These authors have the highest percentage of membership in community #3 of the ACA graph, which is shared among authors of [4] and [6]. None of these authors were cited in [4] or [6]. We showed our findings to domain experts and received validation from them that we had indeed found the relevant researchers.
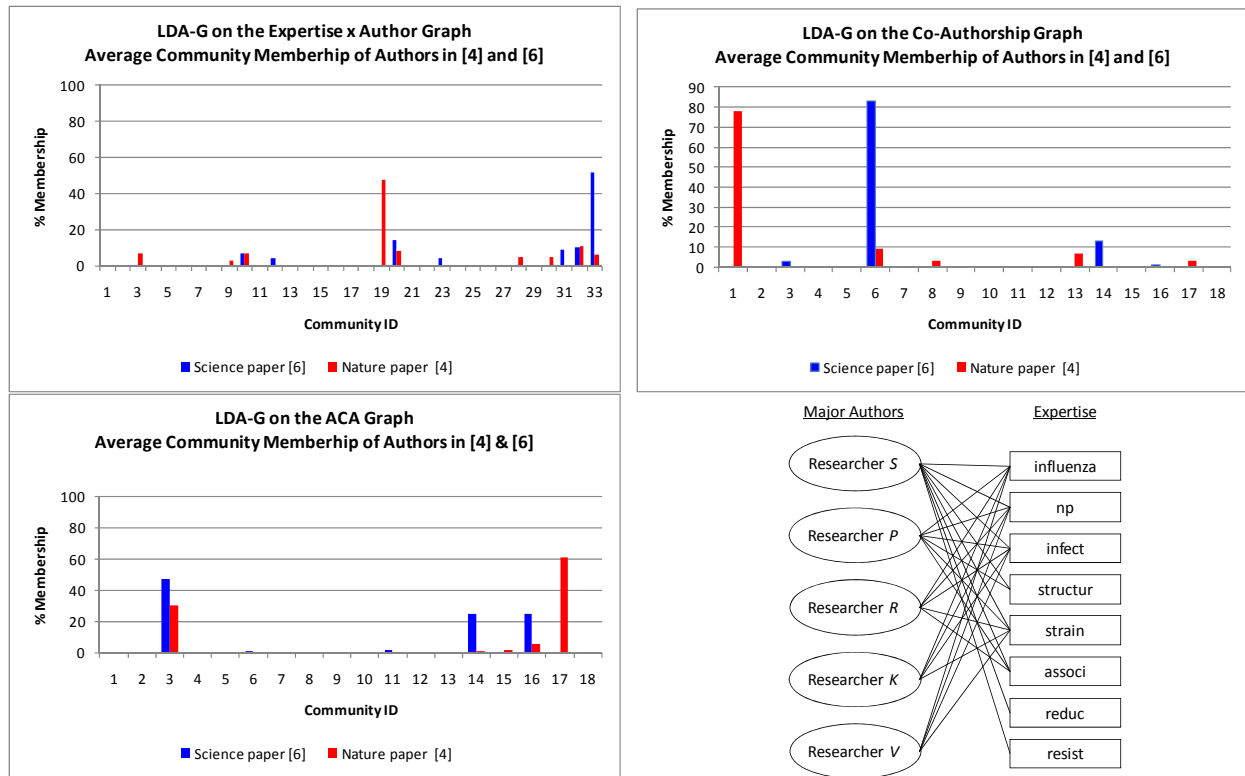

Figure 2. The top row and bottom left plots show the average community membership of authors of [4] and [6]. Only in the ACA graph do we find a common community (#3) with significant overlap between the authors of [4] and [6]. The lower left plot depicts the authors with the highest percentage of membership in community #3 of the ACA graph. These five authors and the authors of [4] and [6] share similar expertise and have topologically similar co-authorship neighborhoods.

Figure 3 depicts the overlap in the expertise terms for authors of [4] and [6]. 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.
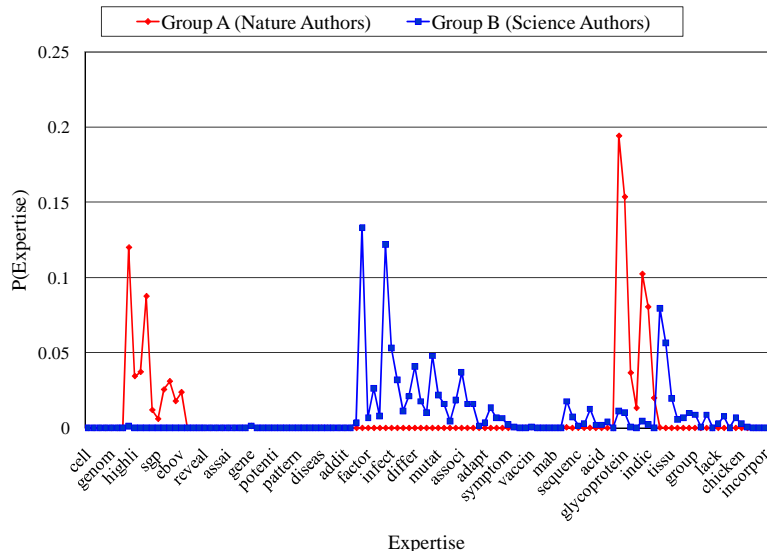
Figure 3. LDA-G's qualitative results on the expertise-by-author graph. Plot shows the probability of expertise terms for major author groups of [4] in red and [6] in blue. Even though both papers are on reconstruction of the 1918 flu virus the authors' expertise terms does not overlap as much as expected.

LDA-G is able to effectively factor out a graph's community structure. Figure 4 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.

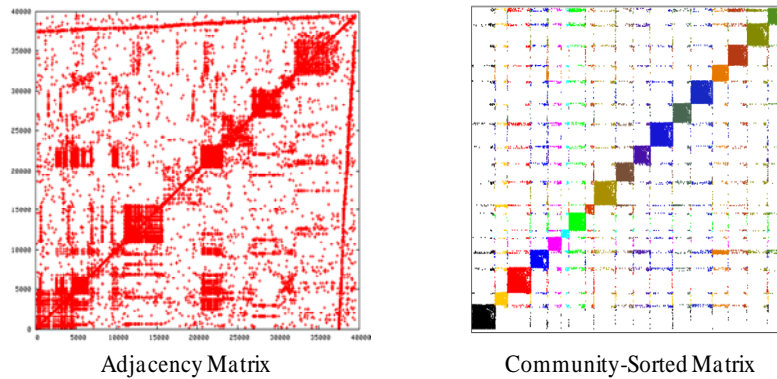Adjacency Matrix   Community-Sorted Matrix

Figure 4. The ACA Graph: its adjacency matrix and its community-sorted matrix.
LDA-G discovers 18 well-separated communities.

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 1 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 discover 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.

Table 1. AUC values on link prediction averaged over 5 trials. (Default value is 0.5.)

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

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.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR* 3:993-1022, 2003.

[2] K. Henderson and T. Eliassi-Rad. Applying latent Dirichlet allocation to group discovery in large graphs. In *ACM SAC* 2009:1456-1461.

[3] K. Henderson, T. Eliassi-Rad, S. Papadimitriou, and C. Faloutsos. Best of both worlds: A scalable hybrid community discovery algorithm. Lawrence Livermore National Laboratory. Technical Report LLNL-TR-414387, July 2009.

[4] D. Kobasa, S. M. Jones, K. Shinya, J. C. Kash, J. Copps, H. Ebihara, Y. Hatta, J. H. Kim, P. Halfmann, M. Hatta, F. Feldmann, J. B. Alimonti, L. Fernando, Y. Li, M. G. Katze, H. Feldmann, and Y. Kawaoka. Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. *Nature* 445(7125), 2007:319-23.

[5] H. Li, Z. Nie, W.-C. Lee, C. L. Giles, and J.-R. Wen. Scalable community discovery on textual data with relations, In *ACM CIKM* 2008:1203-1212.

[6] T. M. Tumpey, C. F. Basler, P. V. Aguilar, H. Zeng, A. Solórzano, D. E. Swayne, N. J. Cox, J. M. Katz, J. K. Taubenberger, P. Palese, and A. García-Sastre. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* 310(5745), 2005:77-80.

[7] H. Zhang, C. L. Giles, H. C. Foley, and J. Yen. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In *AAAI* 2007:663-668.

[8] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An LDA-based community structure discovery approach for large-scale social networks. In *IEEE ISI* 2007: 200-207.

[9] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. Probabilistic models for discovering e-communities, In *WWW* 2006:173-182.