Community Detection and Characteristics Analysis of Academic Network

Qi Zeng, Yao Xiao, Shanshan Xu, Zhefei Yu
(Dated: December 9, 2014)

Clustering and categorization of academic network, especially collaboration network, is always an interesting question, in that it can provide many insight into academic fields. While traditional methods, such as spectral clustering, are based on the assumption that each node can be assigned to only one clustering, more and more observations suggest that communities in academic network probably overlap with each other, or even nested under each other. This motivates us to apply a novel overlapping community detection algorithm on a large collaboration networks we obtain from arXiv with ground-truths. In this paper, we will examine the community detection on our collaboration network for deep insight into the internal structure that was hard to see in the past. By studying leadership characteristics of collaboration network and how leaders distribute among overlapping communities in various years, we not only get more insight into the structure of collaboration network, but find some general disagreement between detected communities and ground-truth as well, which may guide more future model building work.

I. INTRODUCTION

Information encoded inside academic network is always of great interest, especially for people in academics. A typical example of academic network is the paper citation networks. By exploring citation network internal structure, people are especially interested in the classification and ranking of papers inside large database. Another example is the collaboration network of authors, where co-authors are automatically linked together. A detailed study on such collaboration network can not only reveal relation between scholars in academics, but also provide useful guide for people to find resources of interest.

Extensive studies have been done in past years on how to study resolved structure of academic network and basically all of them involve community detection (or clustering). Basically, the task is trying to assign a group label to nodes (either paper or scholar). At early stage, people tend to do an exclusive clustering, which assign an unique group label to each node. Various algorithms are proposed and basically they are trying to identify the dense region and then dissect the network at linkes between dense regions.

However, in recent years, people realize that the communities inside network may, by nature, overlap with each other. It is even possible that one community is completely nested under a larger community. This turns out to be a big challenge. As revealed by previous studies and our observation that will be shown later, academic network, for example collaboration network, is a typical "heavily overlapped" network. In addition, collaboration network is usually big, and may accumulate with time. This requires us to use such an algorithm that can not only detect overlapped communities, but scale with number of nodes as well.

Furthermore, we are interested in identifying those "star" scholars and those hot fields based on academic network. It is also a open question whether BIGCLAM can successfully reproduce the same result as ground-truth in describing the distribution of leaders among communities. By using the ground-truth community information and sophisticated Page Rank algorithm, we are able to study the connection between leadership of a network and the hot fields, as well as how this connection change over the time. By comparing such feature between ground-truth and detected community obtained by novel overlapping community detection algorithm above, we can not only get a more detailed evaluation on how well the community detection works, but gain some insight in the future model improvement as well.

II. RELAVANT PRIOR WORK REVIEW

As shown in the first study of scientific collaboration network [1], scientific collaboration network is more like small-world model with high clustering coefficient and small distance. In particular, it is observed in [1] that degree distribution is more like a power-law with exponential cut-off.

There has been a lot of community detection algorithms developed on social networks [2]. Among them, [3] is a seminal work with application on real collaboration networks. In [3], communities are constructed by progressively removing edges from the original graph based on the defined edges (geodesic) betweenness. Other weights such as random-walk betweenness and current-flow betweenness are defined in later [4]. This algorithm suffers several problems. Computing edge betweenness repeatly is expensive. More importantly, it does not allow overlapping between communities.

Overlapping community detection algorithms are reviewed in [5]. These algorithms, though assuming overlap, bears a hidden assumption that nodes with high degree tend to be within a community instead of being shared by several communities. However, [6] studies various realistic network, including collaboration network, with ground-truth and finds that the more number of communities a pair of node share, the more possible they are connected to each other. Based on this novel observation, a new overlapped community detection is designed in [7]. It use the observation found in [6] and do a likelihood fit over the affiliation matrix. In addition, the algorithm is optimized so that it can scale with number of nodes to millions, which is essential to our study as shown later.

III. MODELS AND ALGORITHMS

The main algorithm we used for the overlapping community detection is BIGCLAM (Cluster Affiliation Model for Big Networks) developed in [7].

This algorithm assumes the existence of a latent interaction of non-negative strength $X_{uv}^{(c)}$ between nodes u, v that are in the same community c. Then the total amount of interaction X_{uv} between nodes u, v is the summation of $X_{uv}^{(c)}$ from all communities. Further, for each community c, $X_{uv}^{(c)}$ is assumed to be an independent Poisson distribution with mean $F_{uc}F_{vc}$, where F_{uc} is the measure of the connection strength between the node u and the community c. As a result, X_{uv} is also a Poisson distribution with mean $\sum_c F_{uc}F_{vc} \equiv F_u \cdot F_v^T$.

Now for the observed network, there is an edge between a pair of nodes u and v if $X_{uv} > 0$. Therefore, for the observed network, edge creation between node u and v has probability

$$p(u, v) = p(X_{uv} > 0) = 1 - P(X_{uv} = 0) = 1 - e^{F_u \cdot F_v^T}. \tag{1}$$

The overlapping community detection problem has two steps:

• Given the observed network G, we learn the parameter F by maximizing the log-likelihood:

$$\hat{F} = \arg\max_{F \geq 0} l(F), \tag{2}$$

where the log-likelihood $l(F)$ is

$$l(F) = \log P(G|F) = \sum_{(u,v)\in E} \log(1 - e^{F_u \cdot F_v^T}) - \sum_{(u,v)\notin E} F_u \cdot F_v^T. \tag{3}$$

The optimization problem of (2) can be solved by a block coordinate gradient ascent algorithm, where F_u is updated for each u with other F_v fixed. The subproblem

$$\arg\max_{F_{uc} \geq 0} \left[\sum_{v \in \mathcal{N}(u)} \log(1 - e^{F_u \cdot F_v^T}) - \sum_{v \notin \mathcal{N}(u)} F_u \cdot F_v^T \right] \tag{4}$$

is a convex optimization problem and can be solved by projected gradient ascent method.

To make comparison, with the performance of the BIGCLAM, we also used the CPM (clique percolation) as the baseline

• After we learn $\hat{F}$, each node u is classified into the community c if $\hat{F}_{uc}$ is larger than some threshold δ.

To evaluate the performance of BIGCLAM, we also run CPM (Clique Percolation Method) [8], another algorithm for overlapping community detection, as the baseline. In this algorithm, a (k-clique-)community is defined as a a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing k-1 nodes). The algorithm has three steps:

• Find maximal cliques (cliques that can't be extended) by BronKerbosch algorithm.

• Create clique overlap matrix. Each row (and column) represents a clique and the matrix elements are equal to the number of common nodes between the corresponding two cliques.

• Threshold the matrix at value $k - 1$. Communities are the connected components of the thresholded matrix.

We expect that BIGCLAM has better performance than CPM since CPM fails to detect dense overlaps.

IV. DATA COLLECTION

Our data is taken from arXiv, through arXiv Bulk Data Access. Each item in data is metadata of each paper, including paper ID, author names, categories(an internal label in arXiv representing various disciplines), dates, and abstract. These metadata are downloaded through Open Archives Initiative (OAI). Then, since we are mostly interested in collaboration network between authors, we build an edge between two authors if they ever co-author a paper.

On the ground-truth side, for each author, we count how many papers he has under each category. Then for each category, we get a weight by dividing such count by total number of papers he ever has:

$$w_i = \frac{N_{field-i}}{N_{total}} \tag{5}$$

Note that papers might have several category tag so the sum of this weight over all categories is not necessarily unity. Then we define a truth-level threshold "truth-threshold", that the author belongs to this category if its weight in this category is larger than "truth-threshold". If the threshold is low (like 0), the categories an author belongs to might be too many since even some occasional submission in a non-related field for fun will also count. But we also do not want to set the threshold too tight, otherwise there will be too little overlap. Eventually we set the threshold to be 0.1

We obtain in total three massive samples in addition to a small sample. The three massive samples are respectively taken from "condensed matter physics", "astrophysics" and "computer science" three large categories. They are chosen since arXiv support further detailed categories under the parent discipline so that our ground-truth is meaningful. All data are taken between 2007-01-01 and 2014-01-01. Note that, though each dataset is taken under a parent discipline (like "condensed matter physics"), the categories are not limited within sub-field of condensed matter physics due to the nature of inter-discipline of some papers. As a matter of fact, there are in total 161 different categories for all data taken under "condensed matter physics", varying from tags in physics, math to tags in computer science.

In all study below, we only use "condensed matter physics" sample since it is more easy to interpret the result based on our background. However, readers may take our other samples in "astrophysics" and "computer science" for further study. There is an additional small sample, which is taken from "condensed matter physics" between 2013-05-01 and 2013-06-01. This sample is small (with only thousands of nodes) so it is only used for testing and debugging.

A visualization by Gephi on small sample can be found in Figure 1

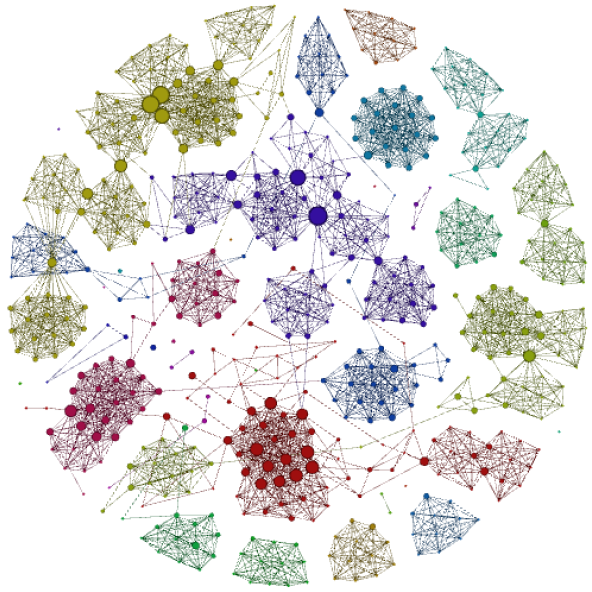

FIG. 1: Visualization by Gephi. The Clustering in Visualization is done by Gephi.

V. RESULTS AND FINDINGS

We organize our results and findings in the following way. An overview of general features of our data will be examined. After checking whether our data is consistent with the underlying assumption made by BIGCLAM algorithm, we begin to apply BIGCLAM algorithm, as well as baseline Clique Percolation Method on our data. Then, we will first evaluate performance of these two methods by using conventional metrics, such as average F1 score etc.

After the algorithm and its performance is established, we begin to study the leadership property in our network. After we define properly the "leader" of a network, we will study how they distribute by using ground truth information, as well as how it evolves since 2007 to 2014 year by year. Later, a similar study, but using detected community, will be shown and tested if they are consistent with our observation from ground-truth. As we will show, such comparison eventually reveal some general disagreement between models behind BIGCLAM algorithm and ground-truth.

A. Network Overview

The sample we use for this part is a combination of all "condensed matter physics" network from 2007-01-01 to 2014-01-01. This collaboration network has 140685 nodes and 744750 edges. There are 2173 nodes with 0 degrees, which are thrown away since they are completely isolated points. We then use largest connected component instead of the whole network as our network. For largest connected component, there are 125827 nodes and 724881 edges.

Figure 2 shows the degree distribution of our network. It can be fit using the power-law. This plot has a nice power-law, but only after some point around 7. By that point, the slope is even positive. The tail of distribution is more like a power-law, instead of an exponential cut-off as observed in [1]

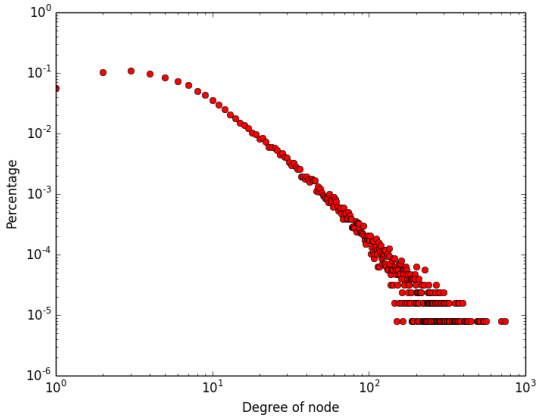

FIG. 2: Degree Distribution

Figure 3 shows the hub and authority score distribution for the network.(figure HubHist.png). Hub and authority scores are usually used for evaluating webpages. For our case, we use this score to evaluate the position of the author in the collaboration network. This will make more sense if we add additional author information like their profiles. However, only based on the property of the network, we can see that the two scores are similar to each other.

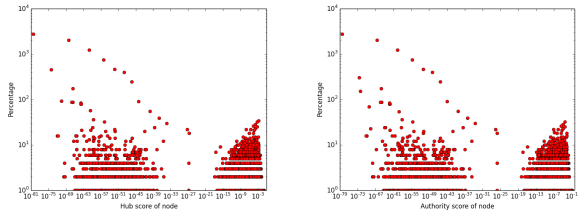

FIG. 3: Hub (left) and Authority (right) Score

Figure 4 (left) shows the distribution of clustering coefficient of the network. The distribution is rather flat, which is a surprising for us.

Figure 4 (right) shows the distribution of number of categories of each author. The familiar power-law distribution suggests that a lot of authors are only in a few categories and only a few of them have published papers in lots of categories.

One should be aware that, for 161 categories we observe under "condensed matter physics", only 9 of them are really the sub-field of "condensed matter physics" with all other categories being "by-product". Part of statistics of categories can

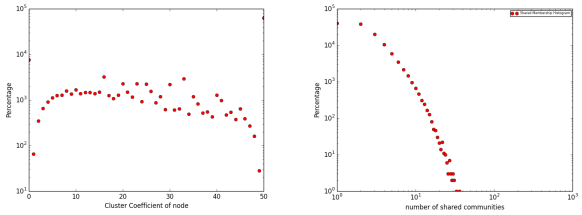

FIG. 4: Clustering Coefficient Distribution and Shared Membership Distribution

be found in Table I. From this table we can see that most items concentrate on a few categories.

category	counts
"cond-mat.mtrl-sci"	53790
"cond-mat.mes-hall"	38489
"cond-mat.str-el"	32224
"cond-mat.stat-mech"	27665
"cond-mat.supr-con"	26123
"cond-mat.other"	18719
"cond-mat.soft"	18369
"cond-mat"	14622
"quant-ph"	13231
"cond-mat.dis-nn"	12929
"cond-mat.quant-gas"	7163
"hep-th"	5258
...	...
"astro-ph"	851
...	...
"cs.DS"	146
...	...

TABLE I: Number of authors in each Categories

Next, we test one of the most important features – edge probability as function of number of shared communities between any pair of the network. This plot will demonstrate whether our data is consistent with the underlying assumption behind BIGCLAM algorithm. In this plot, we ask what is the probability there is an edge between any two nodes given a certain number of common communities they share.

The plots can be found in Figure 5. We overlay curves with three different truth threshold value. All plots are consistent with underlying assumption that the more membership a pair of nodes share, the more possible they will connect with each other.

Another important feature is how edge probability change with truth threshold: as "truth-threshold" increases, the curve will rise faster. Intuitively, as "truth-threshold" goes up, each author will only keep those categories that they mainly work in. As a result, if two authors have more overlap in their main contribution area, the probability of they co-author will increase.

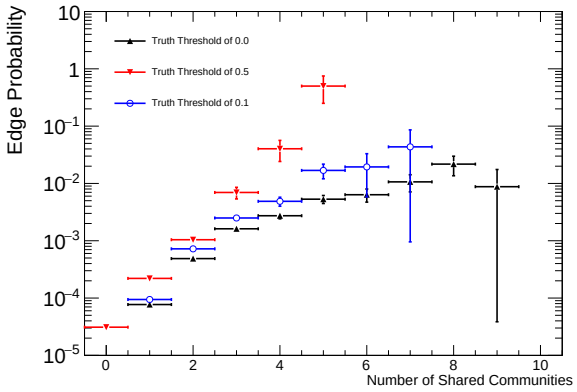

FIG. 5: Edge Probability for Various Truth Threshold

B. Experiment on Collected Data

Having checked edge probability distribution, we can now apply BIGCLAM algorithm on our sample. The implementation is already done in SNAP. The sample we run is full sample combining from 2007 to 2014, as well as sample of each year respectively. In general, the BIGCLAM is very fast (at most 2.5 minutes for largest sample containing all years data), if the number of communities is given. However, if we let SNAP implementation to automatically determine number of communities, it will take forever to run. Therefore, we eventually decide to set the number of communities as in ground-truth as input.

At the same time, we run baseline algorithm, CPM, for comparison. Though CPM is much slower than BIGCLAM, it can still process year-by-year data in reasonable time. The CPM is not applied on combined data since it will take too much time. The CPM implementation in SNAP has one parameter "k". We scan the value "k" from 2 to 20 and choose the one with number of communities closest to ground-truth value. A comparison of performance will be shown in next part.

To see if using number of ground-truth community is a reasonable choice, we run BIGCLAM on a rather small sample (some random month data), in which BIGCLAM can automatically configure optimal number of communities. The curve of optimization of K can be found in Figure 6 and the optimal value is $K = 130$, which is more than ground-truth value 84 for this sample

C. Conventional Performance Evaluation Metrics

Three conventional performance evaluation metrics are adopted here:

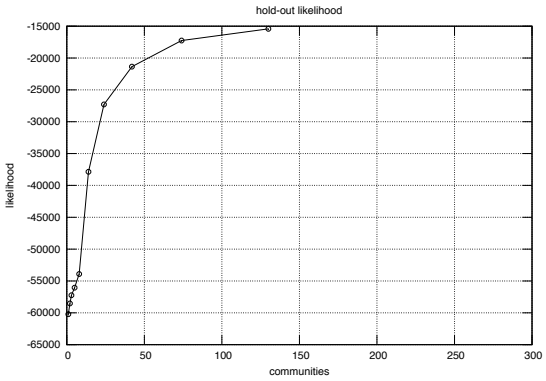

FIG. 6: Likelihood as function of number of communities

- Average F1 Score evaluation, which is defined as

$$\frac{1}{2}\left(\frac{1}{|C^\star|}\sum_{C_i\in C^\star} F1(C_i,\hat{C}_{g(i)}) + \frac{1}{|\hat{C}|}\sum_{\hat{C}_i\in\hat{C}} F1(C_{g'(i)},\hat{C}_i)\right) \tag{6}$$

  where $C^\star$ is truth partition and $\hat{C}$ is detected partition. $g(i) = argmax_j F1(C_i,\hat{C}_j)$ and $g'(i) = argmax_j F1(C_j,\hat{C}_i)$.

- Omega Index, which is used to test accuracy on estimating number of communities that each pair of nodes shares:

$$\frac{1}{|V|^2}\sum_{u,v\in V} 1|C_{uv}| = |\hat{C_{uv}}| \tag{7}$$

- Normalized Mutual Information (NMI), which uses information theory to compare detected communities and ground-truth information. For two partitions Ω and C, NMI is defined as

$$NMI = \frac{I(\Omega;C)}{[H(\Omega+H(C)]/2} \tag{8}$$

  where H is entropy defined as

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N} log\frac{|\omega_k|}{N} \tag{9}$$

  and I is mutual information defined as

$$I(\Omega;C) = \sum_k\sum_j \frac{|\omega_k\cap c_j|}{N} log\frac{N|\omega_k\cap c_j|}{|\omega_k||c_j|} \tag{10}$$

Table II is the tabular showing score of each metric for data from 2007 to 2014 year by year, as well as the comparison with result out of CMP (in bracket). For all tested data, result from BIGCLAM is much better than CPM in average F1 score and NMI metrics. In Omega Index, two algorithm give similar score. These show that BIGCLAM can better detect community in collaboration network than CPM.

year	average F1 score	Omega Index	NMI
2007	0.05598 (0.01008)	0.64960 (0.65006)	0.24323 (0.05494)
2008	0.11338 (0.02299)	0.63526 (0.63429)	0.41711 (0.16523)
2009	0.04920	0.63967	0.24806
2010	0.08555 (0.02380)	0.61531 (0.61428)	0.30048 (0.13015)
2011	0.07803 (0.03517)	0.61518 (0.61476)	0.28038 (0.11235)
2012	0.09387 (0.04927)	0.59354 (0.59253)	0.29931 (0.09679)
2013	0.07003 (0.01980)	0.60078 (0.59975)	0.26491 (0.08859)
2007-2014	0.04474	0.61689	0.22933

TABLE II: Score of Metrics

D. Leadership in Collaboration Network

After community detection, the next thing we want to explore is how leadership distribute inside collaboration network, as well as how it is connected with overlapping communities, which will be studied in this and next part. From academic point of view, the leadership in a collaboration network can indicate people who are active as well as sub-fields that are active.

We define the leadership through Page Rank of each node. Figure 7 show the Page Rank of each node in descendant order. Notice that Y-axis is in log scale. This plot shows only a few people have a very high page rank value, while most people are on the "plateau". Therefore, a practical way to define "leader" would be those nodes with page rank between $[MaxPageRank/2, MaxPageRank]$. It turns out the number of leaders defined in this way is always around 50 so that they are tractable. Also, the exact order of Page Rank among leaders should not be taken too seriously, since, as shown in the Page Rank plot, what really matters is the log-scale value.

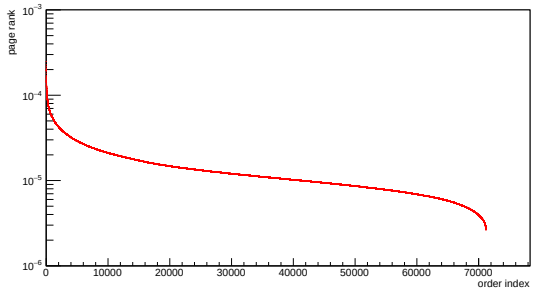

FIG. 7: Page Rank Distribution for some arbitrary year

A distribution of Page Rank from 2007 to 2014 year by year is shown in 8. We find that shape of Page Rank curve is roughly the same over years, though the normalization is different, which strongly depends on the total size of nodes. Generally, the more number of nodes, the lower average Page Rank it has. It also turns out the total number of authors in network is random, except a special year 2009-2010, when network size is significantly larger than other years.

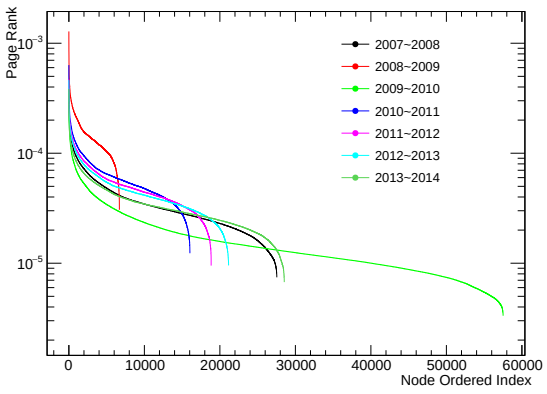

FIG. 8: Page Rank Distribution from 2007 to 2014 Year by Year

E. Leadership Distribution among Communities and How It Changes over Time

As suggested in Table I, the number of nodes per category is highly polarized in ground-truth. Figure 9 shows the number of nodes for each community in descendant order (X-axis is community order index, Y-axis is number of nodes).

Notice that the Y-axis is in log-scale, suggesting the number of papers in each cluster is highly polarized – people tend to aggregate in the hot fields, which are usually defined as the fields with most people involved. Also from Table I, we know that the top 9 categories are actually the real sub-field of "condensed matter physics", while other sub-fields that intersect with condensed matter physics locate at the tail. In general, the non-condensed-matter fields that intersect mostly with condensed matter physics is high energy theoretical physics ("hep-th"), which is not surprising because these two fields sometimes are actually studying the same thing. Astrophysics also have a lot of intersection (actually it can rank top 20 among 100!) with condensed matter physics, which is a bit surprising to me. Another field that has a lot of intersection with condensed matter physics is the "cs.DS"(Computer Science – Data Structure and Algorithm) – considering many physicists who make contribution to network study actually come from condensed matter physics (for exampe, Mark Newman), this is not that surprising.

In the ground-truth plot, we see that as year grows, the number of size in each community is growing as well, which shows more people are contributing to each fields.

Now we focus on the same plot but using detected communities by BIGCLAM (right plot). The trend as function of time is successfully reproduced by BIGCLAM algorithm. However, the shape is very different from ground-truth – detected communities are much less polarized than ground-truth. This implies that BIGCLAM tends to make each communities equally large, which is not true in the ground-truth.

The discussion above is on all nodes in network. Now we only focus on the leaders as defined before. Figure 10 shows the ordered number of leaders in each community from ground-truth. All curves show a very common feature, that

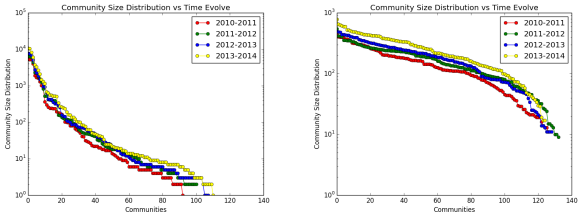

FIG. 9: Ordered Number of Nodes per Community for Ground-truth (left) and Detected Community (right) over years

is there exists four categories which contains most of leaders. Apart from these four categories, other categories can only get very few leaders. This is again, very surprising since we know there are actually in total 9 real sub-fields in condensed matter physics. Furthermore, by looking at ground-truth, these four categories are actually the same in each year from 2007 to 2014: "cond-mat.str-el" (Strongly Correlated Electrons), "cond-mat.mes-hall" (Mesoscale and Nanoscale Physics), "cond-mat.mtrl-sci" (Materials Science) and "cond-mat.supr-con" (Superconductivity). This is consistent with our knowledge in condensed matter physics that these are indeed the four hottest fields.

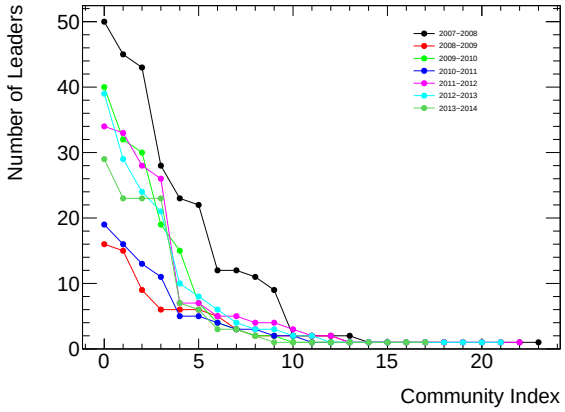

FIG. 10: Ordered Number of Leaders per Community in Ground-truth over Years

However, such sharp feature observed in ground-truth is not reproduced in detected communities, as shown in figure 11. Unlike ground-truth, leaders are more evenly spread out among all communities: most communities can get at least one leader (in ground-truth, only 20 communities can); Although the distribution is still polarized, the drop is not as rapid as ground-truth. Therefore, BIGCLAM algorithm not only tends to "spread out" all members of network among communities, but "spread out" the leaders as well.

Another way to see how leaders distribute among communities is counting number of communities each leader belongs to. Figure 12 shows the ordered number of communities each leader belongs to.

The curves for various years are random and no obvious pattern is observed. Notice that the shape strongly depends on

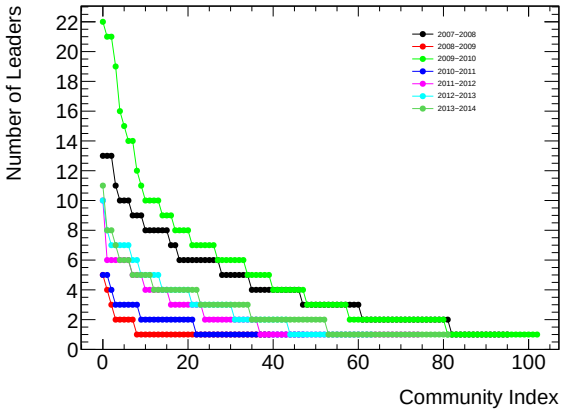

FIG. 11: Ordered Number of Leaders per Community in Detected Communities over Years

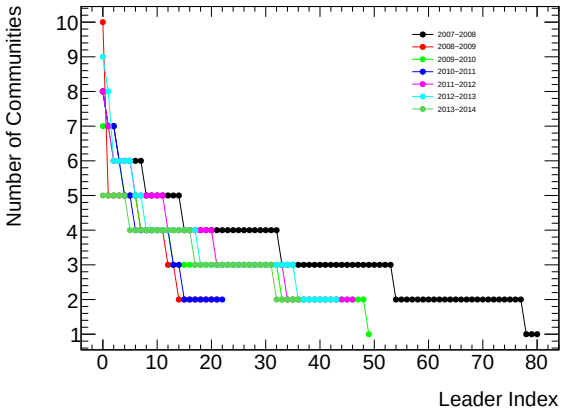

FIG. 12: Ordered Number of Communities per Leader in Ground-truth over Years

number of leaders – after re-scaling X-axis, curves among all years are more or less the same. It turns out there is also a lot of polarization among leaders – only very few leaders belong to more than 5 categories, while most leaders only contribute to one or two fields. More specifically, taking 2007-2008 year range as example, 15% leaders have at least 5 categories (half of maximum), while 61.25% leaders only have at most 3 categories (one third of maximum). This gives us such a picture that there are one or two leaders in the network who have vast connection with many people in other fields, while those people (some of them are leaders as well) only focus on their own fields.

In the detected communities by BIGCLAM, as shown in Figure 13, such polarization is also reproduced to some extend. Since the leaders are defined by Page Rank, which is independent of community detection, we are able to do direct comparison year by year. It turns out leaders are distributed among communities in a much less spread out way than ground-truth. Taking 2007-2008 year as example again, there are 17.5% leaders with at least 9 communities (half of maximal value), and 75% leaders with at most 5 communities (one third of maximal value).

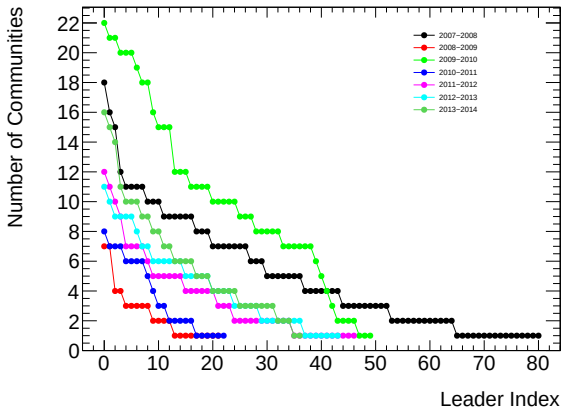

FIG. 13: Ordered Number of Communities per Leader in Detected Communities over Years

So far we have seen how leaders distribute in communities and there is a general disagreement between ground-truth and BIGCLAM algorithm, that ground-truth is more polarized than BIGCLAM. This could come from the assumption behind BIGCLAM: those nodes with high degree tends to be at the overlapping region among all communities. Following this assumption, leaders at the boundary between communities will all tend to have a high number of associated communities. Similarly, if there is a lot of overlapping between communities, most communities will tend to have many leaders.

But ground-truth tells us a slightly different story: Assumption behind BIGCLAM is still correct, but not every node with high degree is the same. Only a few high-degree node is shared by huge amount of communities, while most other high-degree nodes are only shared by 2 or 3 communities. Such observation might be helpful in future improvement of BIGCLAM algorithm.

The last thing about our collaboration network is how the leaders change over year by year. Apparently, from 2007 to 2014, the leadership list will change year by year. In particular, we are interested in two things: 1) People who are always in the leadership from 2007 to 2014; 2) Categories that are always owned by leaders.

For the first question, the answer strongly depends on year range specified by users. For example from 2007 to 2014, there are four authors who are always in the leadership: H. Eisaki, S. Das Sarma, F.M. Peeters and P.C. Canfield. From background knowledge in condensed matter physics, this result makes sense to us. For example, Professor S. Das Sarma is one of the top scholar in theoretical condensed matter physics. Furthermore, by observing the name, there are also many very famous physicists (for example, Professor Shoucheng Zhang of Stanford) appearing at some year but disappearing in next year. Therefore, it might be of more interest to study the variation on a year-by-year base.

For the second question, there are in total 8 categories owned by leaders from 2007 to 2014. These 8 categories span

all different sub-fields of condensed matter physics, which makes sense to us. Similarly, in each year, there will be many non-condense-matter-physics field in this list but usually they will disappear in the next year. A study on such variation on a year-by-year base might be more interesting

VI. CONCLUSION

In this paper, we collect and build collaboration network in condensed matter physics in arXiv from 2007 to 2014. Then we apply BIGCLAM community detection algorithm on our data on a year-by-year base. Then, by studying how leadership distributes between communities, we not only gain more insight into structure of collaboration network, but observe some general disagreement between ground-truth and detected community, which reveals some shortcome of BIGCLAM algorithm:

- The underlying assumption of BIGCLAM, that the more number of communities a pair of node shares, the more possible they will connect with each other, is still sound in our data
- This assumption implies that high-degree node tends to be at the overlapping region between communities
- However, not all high-degree node is as important as each other. Observation from ground-truth suggests these high-degree nodes could also be highly polarized. Only very few of them is at the boundary of huge number of communities, while most of leaders have only 2 or 3 communities
- The picture is: very few leaders are connecting over all main communities, while other leaders are only connecting a few closeby communities.
- No obvious pattern is observed in terms of how leaders distribute between communities change over years.

[1] M. E. J. Newman, The structure of scientific collaboration networks, PNAS, vol. 98 no. 2 404-409, 2001.

[2] S. Fortunato, Community detection in graphs, Physics Reports 486 (3), 75-174, 2010.

[3] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 8271-8276, 2002.

[4] M. E. J. Newman, and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.

[5] Xie, Jierui, Stephen Kelley, and Boleslaw K. Szymanski. "Overlapping community detection in networks: The state-of-the-art and comparative study." ACM Computing Surveys (CSUR) 45.4 (2013): 43.

[6] Yang, Jaewon, and Jure Leskovec. "Structure and overlaps of communities in networks." arXiv preprint arXiv:1205.6228 (2012).

[7] Yang, Jaewon, and Jure Leskovec. "Overlapping community detection at scale: a nonnegative matrix factorization approach." Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013.

[8] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435.7043 (2005): 814-818.