Symmetrizations for Clustering Directed Graphs

Venu Satuluri
Dept. of Computer Science and Engineering
The Ohio State University
satuluri@cse.ohio-state.edu

Srinivasan Parthasarathy
Dept. of Computer Science and Engineering
The Ohio State University
srini@cse.ohio-state.edu

ABSTRACT

Graph clustering has generally concerned itself with clustering undirected graphs; however the graphs from a number of important domains are essentially directed, e.g. networks of web pages, research papers and Twitter users. This paper investigates various ways of symmetrizing a directed graph into an undirected graph so that previous work on clustering undirected graphs may subsequently be leveraged. Recent work on clustering directed graphs has looked at generalizing objective functions such as conductance to directed graphs and minimizing such objective functions using spectral methods (such approaches are also shown to fit in our graph symmetrization framework). We show that more meaningful clusters (as measured by an external ground truth criterion) can be obtained by symmetrizing the graph using measures that capture in- and out-link similarity, such as bibliographic coupling and co-citation strength. However, direct application of these similarity measures to modern large-scale power-law networks is problematic because of the presence of hub nodes, which become connected to the vast majority of the network in the transformed undirected graph. We carefully analyze this problem and propose a Degree-discounted similarity measure which is much more suitable for large-scale networks. We show extensive empirical validation.

1. INTRODUCTION

A number of complex systems and applications can be modeled in the form of a relationship graph or network. Examples abound ranging from Protein Interaction Networks to Twitter networks, from Citation networks to technological networks such as the hyperlinked structure on the World Wide Web. Analyzing such networks can yield important insights about the domain problem in question. A common analysis tool here is to discover the community or cluster structure within such networks.

Directed graphs are essential in domains where relationships between the objects may not be recriprocal i.e., there may be an implicit or explicit notion of directionality in the context of the complex system being modeled. Most of the work to date on community discovery or clustering of graphs has targeted undirected networks and very little has focused on the thorny issue of community discovery in directed networks as noted in a recent survey on the topic[6].

A major challenge is that the nature of relationships captured by the edges in directed graphs is fundamentally different from that for undirected graphs. Consider a citation network where an edge exists from paper i to j if i cites j. Now i may be a paper on databases that cites an important result from the algorithmic literature (j). Our point is that paper i need not necessarily be similar to paper j. A common approach to handle directionality is to ignore it – i.e. eliminate directionality from edges and compute communities. In the above example that would not be the appropriate solution. Such a semantics of directionality is also evident in the directed social network of Twitter, where if a person i follows the feed of a person j, it tell us that i thinks the updates of j are interesting but says nothing about the similarity of i and j.

Recent research (summarized in Section 2) has addressed this problem via generalizing objective functions for groups of vertices in undirected graphs to the context of directed graphs. Frequently, such approaches have relied on spectral clustering in one form or another, but spectral clustering unfortunately does not scale very well. In this article, we instead propose to solve the directed graph clustering problem via a two stage approach; in the first stage, the graph is symmetrized in one of several possible ways, and in the second stage, the so-obtained symmetrized graph is clustered using any state-of-the-art (undirected) graph clustering algorithm. Our argument for pursuing a symmetrization approach is that, if there exists a good symmetric similarity measure - and domains which have reasonable cluster structure must possess a coherent notion of similarity - one can use the similarity measure to induce a symmetrized graph suitable for subsequent clustering. Furthermore, we show that the directed spectral clustering efforts of previous researchers can be cast in our two-stage framework, enabling the use of more scalable graph clustering algorithms other than spectral clustering. We also draw together existing work to define Bibliometric symmetrization, which takes into account the number of common in- and out- links between nodes. However, Bibliometric symmetrization does not work well with large-scale power-law like graphs, as it does not handle hub nodes well. Moreover, we also need a suitable similarity measure between nodes in a directed graph. With these desiderata in mind, we propose Degree-discounted symmetrization, where we explicitly account for the influence of hub nodes via a degree-discounting process.

Ours is, to the best of our knowledge, the first comprehensive comparison of different graph symmetrization techniques. We perform evaluation on four real datasets, three of which (Wikipedia, LiveJournal and Flickr) have million plus nodes, and two (Wikipedia,Cora) of which have dependable ground truth for evaluating the resulting clusters. We examine the characteristics of the different symmetrized graphs in terms of their suitableness for subsequent clustering. Our proposed Degree-discounting symmetrization approach achieves a 22% improvement in F scores over a state-of-the-art directed spectral clustering algorithm on the Cora dataset, and furthermore is two orders of magnitude faster. The degree-discounting symmetrization is also shown to enable clustering that is at least 4-5 times faster than other symmetrizations on our large scale datasets, as well as enabling a 12% qualitative improvement in Wikipedia. We also show compelling examples of the clusters that our symmetrization enables recovery of in the Wikipedia dataset; the structure of this cluster conforms with the examples we have used to motivate our symmetrizations early on in the paper.

2. PRIOR WORK

2.1 Normalized cuts for directed graphs

Many popular methods for clustering undirected graphs search for subsets of vertices with low *normalized cut* [9, 13, 16] (or *conductance*[9], which is closely related). The normalized cut of a group of vertices $S \subset V$ is defined as[16, 13]

$$Ncut(S) = \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{i\in S} degree(i)} + \frac{\sum_{i\in S, j\in \bar{S}} A(i,j)}{\sum_{j\in \bar{S}} degree(j)} \quad (1)$$

where A is the (symmetric) adjacency matrix and $\bar{S} = V - S$ is the complement of S. Intuitively, groups with low normalized cut are well connected amongst themselves but are sparsely connected to the rest of the graph.

The connection between random walks and normalized cuts is as follows [13] : $Ncut(S)$ in Equation 1 is the same as the probability that a random walk that is started in the stationary distribution will transition either from a vertex in S to a vertex in $\bar{S}$ or vice-versa, in one step [13]

$$Ncut(S) = \frac{Pr(S \to \bar{S})}{Pr(S)} + \frac{Pr(\bar{S} \to S)}{Pr(\bar{S})} \quad (2)$$

Using the unifying concept of random walks, Equation 2 have been extended to directed graphs by Zhou et. al. [18] and Huang et. al. [8]. Let P be the transition matrix of a random walk on the directed graph, with π being its associated stationary distribution vector (e.g. PageRank vector) satisyfing $\pi P = \pi$. The probability that a random walk started in the stationary distribution traverses a particular directed edge $u \to v$ is given by $\pi(u)P(u,v)$. The $Ncut$ of a cluster S is again the probability of a random walk transitioning from S to the rest of the graph, or from the rest of the graph into S in one step:

$$Ncut_{dir}(S) = \frac{\sum_{i\in S, j\in \bar{S}} \pi(i)P(i,j)}{\sum_{i\in S} \pi(i)} + \frac{\sum_{j\in \bar{S}, i\in S} \pi(j)P(j,i)}{\sum_{j\in \bar{S}} \pi(j)} \quad (3)$$

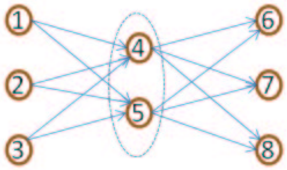


Figure 1: The nodes 4 and 5 form a natural cluster even though they don't link to one another, as they point to the same nodes and are also pointed to by the same nodes.

Meila and Pentney [12] introduce a general class of weighted cut measures on graphs, called $WCut$, parameterized by the vectors T, T' and the matrix A:

$$WCut(S) = \frac{\sum_{i\in S, j\in \bar{S}a} T'(i)A(i,j)}{\sum_{i\in S} T(i)} + \frac{\sum_{j\in \bar{S}, i\in S} T'(j)A(j,i)}{\sum_{j\in \bar{S}} T(j)} \quad (4)$$

Different $NCut$ measures can be recovered from the above definition by plugging in different values for T, T' and A, including the definitions for $NCut$ and $NCut_{dir}$ given above.

All of the above work minimizes these various cut measures via spectral clustering i.e. by post-processing the eigenvectors of suitably defined *Laplacian* matrices. The Laplacian matrix $\mathcal{L}$ for $Ncut_{dir}$, e.g., is given by [18, 8, 3]

$$\mathcal{L} = I - \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P' \Pi^{1/2}}{2} \quad (5)$$

where P is the transition matrix of a random walk, and Π is a diagonal matrix with $diag(P) = \pi$, π being the stationary distribution associated with P.

2.1.1 Drawbacks of normalized cuts for directed graphs

A common drawback of the above line of research is that there exist meaningful clusters which do not necessarily have a low directed normalized cut. The prime examples here are groups of vertices which do not point to one another, but all of which point a common set of vertices (which may belong to a different cluster) We present an idealized example of such situations in Figure 1, where the nodes 4 and 5 can be legitimately seen as belonging to the same cluster, and yet the $Ncut_{dir}$ for such a cluster will be high (the probability that a random walk transitions out of the cluster $\{4, 5\}$ to the rest of the graph, or vice versa, in one step, is very high.) Such situations may be quite common in directed graphs. Consider, for example, a group of websites that belong to competing companies which serve the same market; they may be pointing to a common group of websites outside themselves (and, similarly be pointed at by a common group of websites), but may not point at one another for fear of driving customers to a competitor's website. Another example may be a group of research papers on the same topic which are written within a short span of time and therefore do not cite one another, but cite a common set of prior work and are also in the future cited by the same papers. We present real examples of such clusters in Section 5.5.

Another drawback of the above line of research is the poor scalability as a result of the dependence on spectral clustering (except for Andersen et. al. [1] who use local partitioning algorithms). We further discuss this issue in Section 3.2.

2.2 Bibliographic coupling and co-citation matrices

The *bibliographic coupling* matrix was introduced by Kessler [11], in the field of bibliometrics, for the sake of counting the number of papers that are commonly cited by two scientific documents. It is given by $B = AA^T$, and $B[i,j]$ gives the number of nodes that the nodes i and j both point to in the original directed graph with adjacency matrix A.

$$\begin{aligned} B(i,j) &= \textstyle\sum_k A(i,k)A(j,k) = \sum_k A(i,k)A^T(k,j) \\ B &= AA^T \end{aligned}$$

The *co-citation* matrix was introduced by Small [17], again in the field of bibliometrics. It is given by $C = A^T A$, and $C[i,j]$ gives the number of nodes that commonly point to both i and j in the original directed graph.

3. GRAPH SYMMETRIZATIONS

We adopt a two-stage approach for clustering directed graphs, schematically depicted in Figure 2. In the first stage we transform the directed graph into an undirected graph (i.e. symmetrize the directed graph) using one of different possible symmetrization methods. In the second stage, the undirected graph so obtained is clustered using one of several possible graph clustering algorithms. The advantage of this approach is that it allows a practitioner to employ a graph clustering algorithm of their choice in the second stage. For example, spectral clustering algorithms are typically state-of-the-art quality-wise, but do not scale well as eigenvector computations can be very time-consuming [4]. Under such circumstances, it is useful to be able to plug in a scalable graph clustering algorithm of our own choice, such as Graclus [4], MLR-MCL [15], Metis [10] etc. Note that it is not the objective of this paper to propose a new (undirected) graph clustering algorithm or discuss the strengths and weaknesses of existing ones; all we are saying is that whichever be the suitable graph clustering algorithm, it will fit in our framework.

Of course, the effectiveness of our approach depends crucially on the the symmetrization method. If the symmetrization itself is flawed, even a very good graph clustering algorithm will not be of much use. But do we have reason to believe that an effective symmetrization of the input directed graph is possible? We believe the answer is yes, at least if the domain in question does indeed have some cluster structure. Fundamentally a cluster is a group of objects that are similar to one another and dissimilar to objects not in the cluster. If a domain admits of clusters, this means that there must exist some reasonable similarity measure among the objects in that domain. Since similarity measures are generally symmetric (i.e. $similarity(i,j) = similarity(j,i)$) and positive, defining a notion of similarity for a fixed set of input objects is equivalent to constructing an undirected graph among them, with edges between pairs of objects with non-zero similarity between them and the edge weight equal to the actual value of the similarity. In fact, our proposed degree-discounted symmetrization method can just as validly be thought of as measuring the similarity between pairs of vertices in the input directed graph.

We next discuss various ways to symmetrize a directed graph. In what follows, G will the original directed graph with associated (asymmetric) adjacency matrix A. G_U will be the resulting symmetrized undirected graph with associated adjacency matrix U.

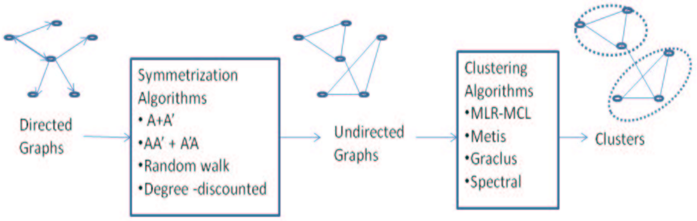


Figure 2: Schematic of our framework

3.1 A+A'

The simplest way to derive an undirected graph from a directed one is via the transformation $U = A + A'$. Note that this is very similar to the even simpler strategy of simply ignoring the directionality of the edges, except that in the case of pairs of nodes with directed edges in both directions, the weight of the edge in the symmetrized graph will be the sum of the weights of the two directed edges. It is important to empiricallly compare this scheme against other symmetrizations since this is the implicit symmetrization used by most algorithms.

The advantage of this method is, of course, its simplicity. On the other hand, this method will fare poorly with situations of the sort depicted in Figure 1; the nodes 4 and 5 will continue to remain unconnected in the symmetrized graph, making it impossible to cluster them together.

3.2 Random walk symmetrization

Is it possible to symmetrize a directed graph G into G_U such that the directed normalized cut of a group of vertices S, $NCut_{dir}(S)$ is equal to the (undirected) normalized cut of the same group of vertices in the symmetrized graph G_U? The answer turns out to be yes.

Let P be the transition matrix of the random walk, π its associated stationary distribution, and Π is the diagonal matrix with π on the diagonal.Let U be the symmetric matrix such that

$$U = \frac{\Pi P + P'\Pi}{2}$$

Gleich [7] showed that for the symmetrized graph G_U with associated adjacency matrix U, the (undirected) Ncut on this graph is equal to the directed Ncut on the original directed graph, for any subset of vertices S. This means that clusters with low directed ncut can be found by clustering the symmetrized graph G_U, and one can use any state-of-the-art graph clustering for finding clusters with low ncut in G_U, instead of relying on expensive spectral clustering using the directed Laplacian (given in Eqn 5) as previous researchers have [18, 8].

The matrix P can be obtained easily enough by normalizing the rows of input adjacency matrix A, and the stationary distribution π can be obtained via power iterations. owever, the clusters obtained by clustering G_U will still be subject to the same drawbacks that we pointed out in Section 2.1.1. Also note that this symmetrization leads to the exact same set of edges as $A + A'$, since P and P' have the same non-zero structure as A and A' and Π is a diagonal matrix. The actual weights on the edges will, of course, be different for the two methods.

3.3 Bibliometric symmetrization

One desideratum of the symmetrized graph is that edges should be present between nodes that share similar (in- or out-) links, and edges should be absent between nodes in the absence of shared (in- or out-)links. Both $A + A'$ and Random walk symmetrization fail in this regard as they retain the exact same set of edges as in the original graph; only the directionality is dropped and, in the case of the Random walk symmetrization, weights are added to the existing edges.

The bibliographic coupling matrix (AA') and the co-citation strength matrix ($A'A$) are both symmetric matrices that help us satisfy this desideratum. Recall that AA' measures the number of common out-links between each pair of nodes, where as $A'A$ measures the number of common in-links. As there does not seem to be any obvious reason for leaving out either in-links or out-links, it is natural to take the sum of both matrices so as to account for both. In this case $U = AA' + A'A$, and we refer to this as bibliometric symmetrization.

Note that not only will new edges be added to the resulting symmetrized graph, but existing edges may also be *removed*. This is actually beneficial because in a directed graph, the mere presence of an edge does not actually indicate affinity between the two vertices, as we have argued in Section 1. However, if the user does not wish for the removal of edges that exist in the original graph, this can be accomplished by simply setting $A := A + I$ prior to the symmetrization.

3.4 Degree-discounted symmetrization

As a consequence of the well-known fact that the degree distributions of many real world graphs follow a power law distribution [5, 2], nodes with degrees in the tens as well as in the thousands co-exist in the same graph. (This is true for both in-degrees and out-degrees.) This wide disparity in the degrees of nodes has implications for the Bibliometric symmetrization; nodes with high degrees will share a lot of common (in- or out-) links with other nodes purely by virtue of their higher degrees. This is the motivation for our proposed *Degree-discounted* symmetrization approach, where we take into account the in- and out-degrees of each node in the symmetrization process.

Another motivation for our proposed symmetrization is defining a useful similarity measure between vertices in a directed graph. As noted earlier in Section 3, a meaningful similarity measure will also serve to induce an effective symmetrization of the directed graph; ideally, we want our symmetrized graph to place edges of high weight between nodes of the same cluster and edges of low weight between nodes in different clusters.

How exactly should the degree of nodes enter into the computation of similarity between pairs of nodes in the graph? First we will consider how the computation of out-link similarity (i.e. the bibliographic coupling) should be changed to incorporate the degrees of nodes.

Consider the following two scenarios (see Figure 3(a)):

1. Nodes i and j both point to the node h, which has in-coming edges from many nodes apart from i and j. In other words, the in-degree of h, $D_i(h)$ is high.

2. Nodes i and j both point to the node k, but which has in-coming edges only from a few other nodes apart from i and j.

Intuition suggests that case 1 above is a more frequent (hence less informative) event than case 2, and hence the former event should contribute less towards the similarity between i and j than the latter event. In other words, *when two nodes i and j commonly point to a third node, say l, the contribution of this event to the similarity between i and j should be inversely related to the in-degree of l.*

Next we consider how the degree of two nodes should factor into the similarity computation of those two nodes themselves. Figure 3(b) illustrates the intuition here: sharing a common out-link k counts for less when one of the two nodes that are doing the sharing is a node with many out-links. In other words, *the out-link similarity between i and j should be inversely related to the out-degrees of i and j.*

We have determined qualitatively how we should take the in- and the out-degrees of the nodes into account, but the exact form of the relationship remains to be specified. We have found experimentally that discounting the similarity by the square root of the degree yields the best results; making the similarity inversely proportional to the degree itself turned out to be an excessive penalty.

Pulling all of the above insights together, we modify the expression for the bibliographic coupling or the out-link similarity $B_d(i,j)$ between the nodes i and j as follows: (recall that D_o is the diagonal matrix of out-degrees, and $D_o(i)$ is short-hand for $D_o(i,i)$. Similarly D_i is the diagonal matrix of in-degrees.)

$$\begin{aligned} B_d(i,j) &= \frac{1}{\sqrt{D_o(i)}\sqrt{D_o(j)}} \sum_k \frac{A(i,k)A(j,k)}{\sqrt{D_i(k)}} \\ &= \frac{1}{\sqrt{D_o(i)}\sqrt{D_o(j)}} \sum_k \frac{A(i,k)A^T(k,j)}{\sqrt{D_i(k)}} \end{aligned}$$

We refer to this as the *degree-discounted bibliographic coupling* and denote it by B_d. Note that the above expression is symmetric in i and j, as any meaningful similarity measure should be.

It can be verified that the entire matrix B_d with its (i,j) entries specified as above can be expressed as:

$$B_d = D_o^{-1/2} A D_i^{-1/2} A^T D_o^{-1/2} \tag{6}$$

Our modification for the co-citation (in-link similarity) matrix is exactly analogous to the above discussion; we proceed to directly give the expression for the matrix C_d containing the degree-discounted co-citation or in-link similarities between all pairs of nodes.

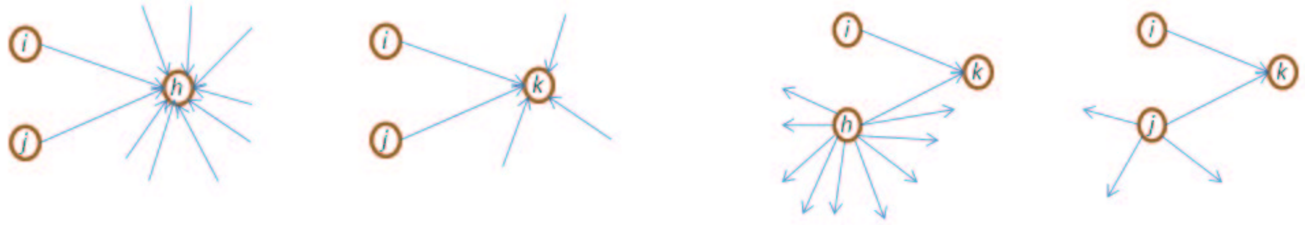
$$C_d = D_i^{-1/2} A^T D_o^{-1/2} A D_i^{-1/2} \tag{7}$$

The final degree discounted similarity matrix U_d is simply the sum of B_d and C_d.

$$\begin{aligned} U_d &= B_d + C_d \\ &= D_o^{-1/2} A D_i^{-1/2} A^T D_o^{-1/2} + D_i^{-1/2} A^T D_o^{-1/2} A D_i^{-1/2} \end{aligned}$$

3.5 Pruning the symmetrized matrix

One of the main advantages of Degree-discounted symmetrization over Bibliometric symmetriation ($AA' + A'A$) is that it is much easier to prune the resulting matrix. $AA' + A'A$ and the Degree-discounted similarity matrix U_d share the same non-zero structure, but the actual values are, of course, different. For big real world graphs, the full similarity matrix has far too many non-zero entries and clustering the entire resulting undirected graph is very time-



(a) If the nodes i and j both point to a hub node h with many incoming edges (**left**), that should contribue lesser to their similarity than if they commonly point to a non-hub node k (**right**)

(b) All else equal, the node i should be less similar to the hub node h which has many out-going edges (**left**) when compared to the non-hub node j (**right**).

Figure 3: Scenarios illustrating the intuition behind degree-discounting.

consuming. For this reason, it is critical that it be possible for us to pick a threshold so as to be able to retain only those entries in the matrix which pertain to sufficiently similar pairs of nodes. However, picking a threshold for $AA^T + A^T A$ can be very hard; as the degrees of nodes are not taken into account, the hub nodes in the graph generate a large number of non-zero entries with high non-zero values (this is because hubs will tend to share a lot of out-links and in-links with a lot of nodes just by virtue of their having high degrees). When we set a high threshold so as to keep the matrix sparse enough to be able to cluster in a reasonable amount of time, many of the rows corresponding to the other nodes become empty. When we lower the threshold in response, the matrix becomes very dense and it becomes impractical to cluster such a dense matrix.

This problem is considerably reduced when applying Degree-discounted symmetrization. This is because the matrix entries involving hub nodes no longer are the largest; this lets us choose a threshold such that when we retain only matrix entries above the threshold, we have a matrix that is sufficiently sparse and at the same time covers the majority of nodes in the graph.

We discuss this issue in the context of real datasets in Section 5.1

4. EXPERIMENTAL SETUP

4.1 Datasets

We perform experiments using four real datasets, detailed below. Also see Table 4.1.

1. Wikipedia: This is a directed graph of hyperlinks between Wikipedia articles. We downloaded a snapshot of the entire Wikipedia corpus from the Wikimedia foundation [1] (Jan–2008 version). The corpus has nearly 12 million articles, but a lot of these were insignificant or noisy articles that we removed as follows. First, we retained only those articles with an abstract, which cut the number of articles down to around 2.1 million. We then constructed the directed graph from the hyperlinks among these pages and retained only those nodes with out-degree greater than 15. We finally obtained a directed graph with 1,129,060 nodes and 67,178,092 edges, of which 42.1% are bi-directional.

Pages in Wikipedia are assigned to one or more categories by the editors (visible at the bottom of a page), which we used to prepare ground truth assignments for the pages in our dataset. We removed the many categories that are present in Wikipedia for housekeeping purposes (such as "Articles of low significance", "Mathematicians stubs"). We further removed categories which did not have more than 20 member pages in order to remove insignificant categories. We obtained 17950 categories after this process. Note that these categories are not disjoint, i.e. a page may belong to multiple categories (or none).

2. Cora: This is a directed graph of CS research papers and their citations. It has been collected and shared by Andrew McCallum [2]. Besides just the graph of citations, the papers have also been manually classified into 10 different fields of CS (such as AI, Operating Systems, etc.), with each field further sub-divided to obtain a total of 70 categories at the lowest level. We utilize the classifications at the lowest level (i.e. 70 categories) for the sake of evaluation. This graph consists of 17,604 nodes with 77,171 directed edges. Note that although symmetric links are, strictly speaking, impossible in citation networks (two papers cannot cite one another as one of them will need to have been written before the other), there is still a small percentage (7.7%) of symmetric links in this graph due to noise.

3. Flickr and **4. Livejournal**:These are large scale directed graphs, collected by the Online Social Networks Research group at The Max Planck Institute [14]. The number of nodes and edges for these datasets can be found in Table 4.1. We use these datasets only for scalability evaluation as we do not have ground truth information for evaluating effectiveness of discovered clusters.

4.2 Setup

We compare four different graph symmetrization methods described in Section 3. For Random walk symmetrization, the stationary distribution was calculated with a uniform random teleport probability of 0.05 in all cases. We clustered the symmetrized graphs using MLR-MCL [15], Metis [10] and Graclus [4]. We are able to show the results of Graclus only on the Cora dataset as the program did not finish execution on any of the symmetrized versions of the Wikipedia dataset. Note that the number of output clusters in MLR-MCL can *only be indirectly controlled* via changing some other parameters of the algorithm; for this reason there is a slight variation in the number of clusters output by this algorithm for different symmetrizations.

We also compare against the BestWCut algorithm described by Meila and Pentney [12], but on the Cora dataset alone, as the algorithm did not finish execution on the Wikipedia dataset. It bears emphasis that BestWCut is not a sym-

[1] http://download.wikimedia.org/

[2] http://www.cs.umass.edu/ mccallum/code-data.html

Dataset	Vertices	Edges	Percentage of symmetric links	No. of ground truth categories
Wikipedia	1,129,060	67,178,092	42.1	17950
Cora	17,604	77,171	7.7	70
Flickr	1,861,228	22,613,980	62.4	N.A.
Livejournal	5,284,457	77,402,652	73.4	N.A.

Table 1: Details of the datasets

metrization method. The directed spectral clustering of Zhou et. al. [18] did not finish execution on any of our datasets.

All the experiments were performed on a dual core machine (Dual 250 Opteron) with 2.4GHz of processor speed and 8GB of main memory. However, the programs were single-threaded so only core was utilized. The software for each of the undirected graph clustering algorithms as well as BestWCut [12] was obtained from the authors' respective webpages. We implemented the different symmetrization methods in C, using sparse matrix representations.

4.3 Evaluation method

The clustering output by any algorithm was evaluated with respect to the ground truth clustering as follows. Let the output clustering be $\mathcal{C} = \{C_1, C_2, \ldots, C_i, \ldots, C_k\}$. For any output cluster C_i, the precision and recall of this cluster against a given ground truth category, say G_j, are defined as: $Prec(C_i, G_j) = \frac{|C_i \cap G_j|}{|C_i|}$ and $Rec(C_i, G_j) = \frac{|C_i \cap G_j|}{|G_j|}$. The F-measure $F(C_i, G_j)$ is the harmonic mean of the precision and the recall. We match each output cluster C_i with the ground truth cluster G_j for which $F(C_i, G_j)$ is the highest among all ground truth clusters. This is the F-measure that is subsequently associated with this cluster, and is referred to as $F(C_i)$; i.e.

$$F(C_i) = \max_j F(C_i, G_j)$$

The average F-measure of the entire clustering $\mathcal{C}$ is defined as the average of the F-measures of all the clusters, weighted by their sizes (i.e. we compute the micro-averaged F-measure).

$$Avg.F(\mathcal{C}) = \frac{\sum_i |C_i| * F(C_i)}{\sum_i |C_i|}$$

5. RESULTS

5.1 Characteristics of symmetrized graphs

The number of edges in the resulting symmetrized graph for each strategy, the threshold and the number of singletons can be found in Table 2. We do not perform any pruning using a threshold for $A + A'$ and Random Walk symmetrization as the number of edges for both methods is the same as the number of edges in the original graph. The unpruned graphs for Bibliometric and Degree-discounted can be quite dense however (as mentioned in Section 3.5) and so we pick a threshold below which edges are eliminated from the graph, so that the graph is clusterable in a reasonable amount of time. Doing this may result in singletons,as some nodes may have edges only below the chosen threshold. In fact this problem is quite severe for the Bibliometric symmetrization; even though the size of the thresholded graph is comfortably more than the size of the thresholded Degree-discounted graph, there are many more singletons in the former rather than latter. In the Wikipedia dataset, for example, the Bibliometric symmetrization graph has nearly 50% of the nodes as singletons, and the relevant percentage is 95% and 59% for Flickr and Livejournal, all of which are million-plus node graphs.

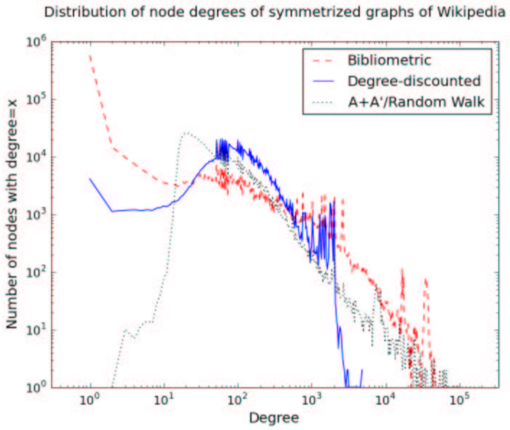


Figure 4: Distributions of node degrees for different symmetrizations of Wikipedia.

We next analyze the distribution of node degrees of each of the symmetrized graphs for Wikipedia, shown in Figure 4. Note that $A + A'$ and Random Walk have the same distributions, as they have the same set of edges. The Degree-discounted method ensures that most nodes have medium degrees in the range of 50-200 (which is about the size of the average cluster), and completely eliminates hub nodes. These properties enable subsequent graph clustering algorithms to perform well. The Bibliometric graph, on the other hand, has many nodes with both very low degrees, as well as many hub nodes, making clustering the resulting graph difficult. The $A + A'$ graph also has more hub nodes than the Degree-discounted graph.

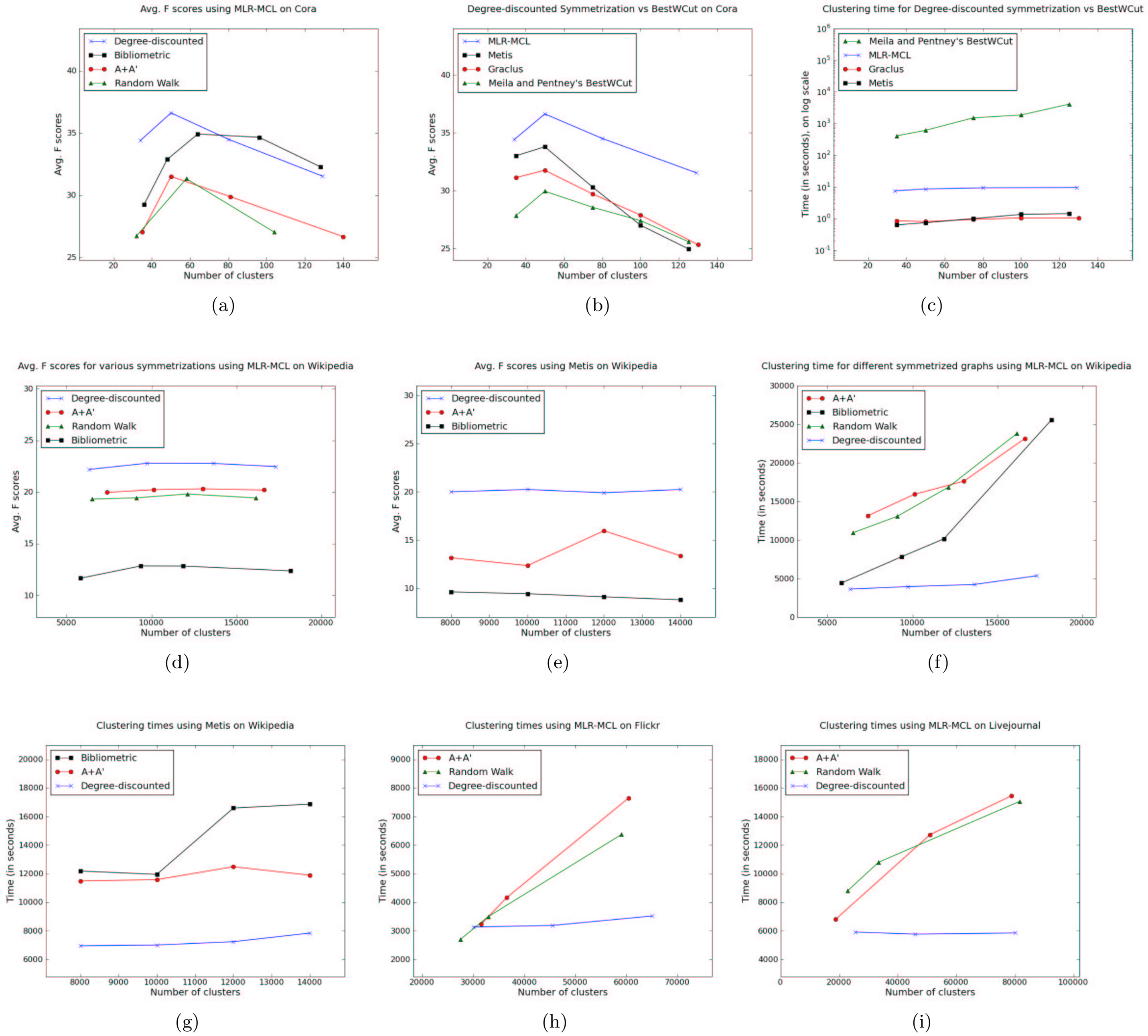
5.2 Results on Cora

Results pertaining to cluster quality as well as clustering time on the Cora dataset are shown in the graphs (a)-(c) in Figure 5.1.

Figure 5.1 (a) compares the Avg. F scores obtained using MLR-MCL with different symmetrizations. For all symmetrizations, the performance reaches a peak at 50-70 clusters, which is close to the true number of clusters (70). With fewer clusters, the precision is adversely impacted, while a greater number of clusters affects the recall. Degree-discounted symmetrization on the whole yields better F scores than the other methods, and also achieves the best overall F-value of 36.62. Bibliometric symmetrization also yields good F-scores with a peak of 34.92, and marginally improves on Degree-discounted for higher number of clusters. $A + A'$ and Random walk perform similarly and are relatively poor compared to the other two methods.

Dataset	$A + A^T$ / Random Walk	Bibliometric			Degree-discounted		
	Edges	Edges	Threshold	Num. Singletons	Edges	Threshold	Num. Singletons
Wikipedia	53,017,527	85,035,548	25	542,403 (48%)	80,373,184	0.01	2910
Flickr	15,555,041	79,765,961	20	1,766,230 (95%)	45,167,216	0.01	181,166 (10%)
Cora	74,180	986,444	0	159	986,444	0	159
Livejournal	51,352,001	143,759,001	5	3096729 (59%)	91,624,309	0.025	111087 (2%)

Table 2: Number of edges,pruning thresholds and the number of singleton nodes in the resulting matrices/graphs for various symmetrization strategies.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 5: The graphs (a)-(c) show results on the Cora dataset. (a) compares cluster quality of different symmetrizations using MLR-MCL for varying number of clusters. (b) compares the cluster quality of Degree-discounted symmetrization using MLR-MCL, Graclus and Metis, with Meila and Pentney's BestWCut. (c) compares the clustering times of the same methods, note that the y-axis is on log scale. The graphs (d)-(g) show results on Wikipedia. (d) and (e) show cluster quality of different symmetrizations using MLR-MCL and Metis, respectively. (f) and (g) show clustering times using MLR-MCL and Metis, respectively. The graphs (h) and (i) show clustering times using MLR-MCL on Livejournal and Flickr, respectively.

Figure 5.1 (b) fixes the symmetrization to Degree discounted and compares MLR-MCL, Graclus and Metis with Meila and Pentney's BestWCut [12]. MLR-MCL comfortably outperforms the other graph clustering algorithms, with the same peak F-value of 36.62. The peak F-values of Metis, Graclus and BestWCut are 33.78, 31.75 and 29.94 respectively. It is also noteworthy that Degree-discounted symmetrization combined with either of the 3 algorithms - MLR-MCL, Metis and Graclus - outperforms BestWCut.

Figure 5.1 (c) compares cluster times of MLR-MCL, Graclus and Metis with Degree-discounted symmetrization against the time taken by BestWCut. All three are much faster than BestWCut. The slow performance of BestWCut is because of the need for expensive eigenvector computations, which none of the other three algorithms involve.

5.3 Results on Wikipedia

We next turn to cluster quality and timing results on Wikipedia, depicted in Figure 5.1(d)-(g). In general, this dataset was harder to cluster than the Cora dataset, with an overall peak Avg. F score of 22.79, compared to 36.62 for Cora. Note that we do not have any results from BestWCut [12] on this dataset as it did not finish execution.

Figure 5.1 (d) and (e) compares the Avg. F scores with different symmetrizations using MLR-MCL and Metis. One common trend in both cases, for all symmetrizations, is that there is not much variation in the Avg F. scores with varying number of clusters. Degree-discounted symmetrization yields the best Avg F scores in all instances, with a peak F value of 22.79. $A + A'$ gives the next best results, with a peak F value of 20.31. The performance of Random Walk is slightly worse than $A + A'$ but is otherwise similar. We do not report Metis combined with Random Walk symmetrization as the program crashed when run with this input. Bibliometric performs very poorly, with F scores barely touching 13%. This is mainly because the large number of singletons in the symmetrized graph. This further bolsters our argument that pure Bibliometric symmetrization is particularly ill-suited for large power-law graphs.

Figure 5.1 (f) and (g) show the time to cluster different symmetrizations using MLR-MCL and Metis. We find that both MLR-MCL and Metis execute faster with Degree-discounted, than any of the other symmetrizations. The difference becomes more pronounced with increasing number of clusters; MLR-MCL executes nearly 4.5 to 5 times faster on Degree-discounted as compared to the other symmetrizations in the high clusters range (16000-18000). We believe that the absence of hub nodes (as can be seen in Fig 4), coupled with clearer cluster structures in the Degree-discounted graph explains its better performance. It is also interesting to note that on this dataset MLR-MCL is on average significantly faster (2000s) than Metis on the degree-discounted transformation.

5.3.1 *Varying the prune threshold*

How does the performance of Degree-discounted symmetrization change as we change the pruning threshold i.e. as more or fewer edges are retained in the graph? We experimented with four different thresholds and clustered the resulting graph with MLR-MCL. The obtained Avg F scores as well as times to cluster are given in Table 5.3.1. (Recall that the number of output clusters can only be indirectly controlled in MLR-MCL.) The trends depicted in the table accord very well with our intuition; as we raise the threshold, there are fewer edges in the graph, and there is a gradual drop in the cluster quality, but which is compensated by faster running times. In fact, even with a threshold of 0.025, and having only 60% as many edges as $A + A'$, Degree-discounted still yields an F score of 21.72 (compared to 20.2 for $A + A'$) and clusters in 1039 seconds (compared to nearly 23000 seconds for $A + A'$).

Threshold	No. of edges	Clusters	Avg F score	Time
0.010	80,373,184	17296	22.47	4225
0.015	73,273,127	16657	22.45	3615
0.020	50,801,885	15347	22.27	1912
0.025	37,663,652	16934	21.72	1039

Table 3: Effect of varying pruning threshold

5.4 Results on Livejournal and Flickr

In Figure 5.1(h) and (i), we show clustering times using MLR-MCL on the Livejournal and Flickr datasets. We could not evaluate cluster quality for lack of ground truth data. We do not report results on Bibliometric, since it is clear from the number of singletons for that transformation (see Table 5.1) that it is not viable for such large scale graphs. The trends for these datasets closely mimic the trends in Wikipedia, with Degree-discounted symmetrization once again proving at least two times as fast to cluster as the others at the higher range of the number of clusters.

5.5 A case study of Wikipedia clusters

Why exactly does Degree-discounted symmetrization outperform other methods? We give some intuition on this question using examples of Wikipedia clusters that were successfully extracted through this method but not with the other symmetrizations. Note that these example clusters were recovered by both MLR-MCL as well as Metis and is thus independent of the clustering algorithms.

A typical example is the cluster consisting of the plant species belonging to the genus *Guzmania*. The in-links and out-links of this group is shown in Figure 6. Example cluster consisting of plants belonging to the *Guzmania* family. The first notable fact about this cluster is that *none of the cluster members* links to one another, but they all point to some common pages - e.g. “Poales”, which is the Order containing the *Guzmania* genus; “Ecuador”, which is the country that all of these plants are endemic to; and so on. All group members are commonly pointed to by the *Guzmania* node as well as point to it in return.

Note that this cluster is not an isolated example. Clusters involving lists of objects particularly were found to satisfy a similar pattern to the *Guzmania* cluster. Other examples include *Municipalities in Palencia, Irish cricketers, Lists of birds by country* etc.

These examples provide empirical validation of our hypothesis - laid out in Section 3 and Figure 1 - that in-link and out-link similarity, and not inter-linkage, are the main clues to discovering meaningful clusters in directed graphs.

5.6 Top-weight edges in Wikipedia symmetrizations

We pick the top-weighted edges in the different symmetrizations of Wikipedia to gain a better understanding into their workings. The top 5 edges from Degree-discounted, Bibliometric and Random Walk symmetrizations are shown in

Symmetrization method	Node 1	Node 2	Edge weight
Random walk	Area	Square mile	3354848
	Mile	Square mile	2233110
	Geocode	Geographic coordinate system	1788953
	Degree (angle)	Geographic coordinate system	1766339
	Area	Octagon	1457427
Bibliometric	Area	Population density	2465
	Record label	Music genre	2423
	Population density	Geographic coordinate system	2301
	Square mile	Population density	2129
	Area	Time zone	2120
Degree-discounted	Cyathea	Cyathea (Subgenus Cyathea)	68
	Roman Catholic dioceses in England & Wales	Roman Catholic dioceses in Great Britain	57
	Sepiidae	Sepia (genus)	55
	Szabolcs-Szatmár-Bereg	Szabolcs-Szatmár-Bereg-related topics	53
	Canton of Lizy-sur-Ourcq	Communauté de communes du Pays de l'Ourcq	52

Table 4: Edges with highest weights for different symmetrization methods on the Wikipedia dataset. Note that the edge weights in the rightmost column are normalized by the lowest edge weight in the graph, as the non-normalized weights are incommensurable.

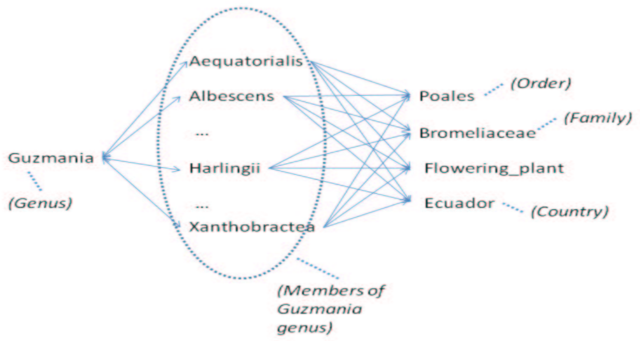


Figure 6: The subgraph of Wikipedia consisting of plant species of the genus *Guzmania* and their in-links and out-links.

Table 4. Bibliometric heavily weights edges involving hub nodes such as 'Area', 'Population density' etc ('Area' has an in-degree of 71,146, e.g.), as expected. Similarly, Random walk heavily weights edges involving nodes with high Page Rank, which also typically tend to be hub nodes. The top-weighted edges of Degree-discounted, on the other hand, involve non-hub nodes with specific meanings; the particular examples listed in Table 4 are almost duplicates of one another.

6. CONCLUSION

In this article, we have investigated the problem of clustering directed graphs through a two-stage process of symmetrizing the directed graph followed by clustering the symmetrized undirected graph using an off-the-shelf graph clustering algorithm. We presented Random Walk and Bibliometric symmetrizations, drawing upon previous work, and based on an analysis of their weaknesses, presented the Degree-discounted symmetrization. We compared the different symmetrizations extensively on large scale real world datasets w.r.t. both quality and scalability, and found that Degree-discounted symmetrization yields significant improvements in both areas. In future work, we would like to investigate the performance of our proposals in large-scale web scenarios involving the possibilities of spam and link fraud. Extending our approaches to bi-partite and multi-partite graphs also seems to be a promising avenue.

7. REFERENCES

[1] R. Andersen, F. R. K. Chung, and K. J. Lang. Local partitioning for directed graphs using pagerank. In *WAW*, pages 166–178, 2007.

[2] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[3] F. Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[4] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):1944–1957, 2007.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, 1999.

[6] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[7] D. Gleich. Hierarchical Directed Spectral Graph Partitioning. 2006.

[8] J. Huang, T. Zhu, and D. Schuurmans. Web communities identification from random walks. *Lecture Notes in Computer Science*, 4213:187, 2006.

[9] R. Kannan, S. Vempala, and A. Veta. On clusterings-good, bad and spectral. In *FOCS '00*, page 367. IEEE Computer Society, 2000.

[10] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20, 1999.

[11] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[12] M. Meila and W. Pentney. Clustering by Weighted Cuts in Directed Graphs. In *SDM*, 2007.

[13] M. Meila and J. Shi. A random walks view of spectral segmentation. In *Artificial Intelligence and Statistics AISTATS*, 2001.

[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[15] V. Satuluri and S. Parthasarathy. Scalable graph clustering using stochastic flows: applications to community discovery. In *KDD '09*, pages 737–746, New York, NY, USA, 2009. ACM.

[16] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[17] H. Small. Co-citation in the scientific literature: A new measure of the relationship between documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[18] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *ICML '05*, pages 1036–1043, 2005.