Noname manuscript No.
(will be inserted by the editor)

The Very Small World of the Well-connected

Xiaolin Shi · Matthew Bonner · Lada A. Adamic · Anna C. Gilbert

Received: date / Accepted: date

Abstract Online networks occupy an increasingly larger position in how we acquire information, how we communicate with one another, and how we disseminate information. Frequently, small sets of vertices dominate various graph and statistical properties of these networks and, because of this, they are relevant for structural analysis and efficient algorithms and engineering. For the web overall, and specifically for social linking in blogs and instant messaging, we provide a principled, rigorous study of the properties, the construction, and the utilization of subsets of special vertices in large online networks. We show that graph synopses defined by the importance of vertices provide small, relatively accurate portraits, independent of the importance measure, of the important vertices in the underlying larger graphs. Furthermore, they can be computed relatively efficiently in real-world networks.

In addition, we study the stability of these graph synopses over time and trace their development in several large dynamic data sets. We show that important vertices are more likely to have longer active life spans than unimportant ones and that the graph synopses consisting of important vertices remain stable over long periods of time after a short period of initial growth.

X. Shi
Department of EECS
University of Michigan
Ann Arbor, MI 48109-2121
E-mail: shixl@umich.edu

M. Bonner
Department of EECS
University of Michigan
Ann Arbor, MI 48109-2121
E-mail: mabonner@umich.edu

L. Adamic
School of Information and
Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI 48109-1107
E-mail: ladamic@umich.edu

A. Gilbert
Department of Mathematics
University of Michigan
Ann Arbor, MI 48109-1043
E-mail: annacg@umich.edu

Keywords E.1 Data Structures, Graphs and networks, G2.2 Graph Theory, Graph algorithms, E.4 Coding and information theory, Data compaction and compression.

1 Introduction

Networks play a crucial role in how we acquire information, how we convey information to one another, and how we interact with other people. On the World Wide Web, we do this through generating and linking content. To study the flow of information, to optimize engineering systems, to design efficient algorithms [5,15, 16], and to investigate social structure and interaction, we study the statistical and graph properties of entire networks, including such features as degree distributions, connectivity, diameter, clustering properties, and evolution of such networks [3,6]. For a variety of online networks, small subsets of vertices dominate various graph and statistical properties. Frequently, these smaller subsets or graph synopses are easier to study and to understand. One might be interested in whether relationships among predominant vertices might be inferred from a small set of vertices. We might also study the "communication" among the most influential political blogs [2] and determine whether information flows directly among them or through intermediate blogs. Despite these examples, there is little principled study of the properties, the construction, and the utilization of subsets of special vertices or edges in large real networks. Such a study is challenging because it is hard to define precisely what is meant by a small version of the graph. Also, it is difficult to evaluate the quality of a compressed graph and we have myriad definitions of important vertices.

We would like a simple, principled approach to graph synopsis for a number of reasons. First, there are a number of online networks in which a synopsis of the graph is sufficient to capture the relevant information we seek. For example, rather than continuously tracking millions of blogs, one may use occasional snapshots of the blogosphere to construct a subgraph of the most "important blogs" according to a desired measure, and crawl, query, and analyze this smaller synopsis. The synopsis will allow us to capture predominant features of the important blogs and, due to its small size, can be stored much more efficiently and even distributed and replicated amongst a number of resource-constrained computers which themselves can execute queries on the content and links.

To build a principled approach to graph synopsis, we start with the definition of predominant vertices and define a precise construction of a graph synopsis from these. Typically, the subset of vertices which capture the graph features are those which are "important." Furthermore, the importance of these vertices is highly skewed—only few of them are of great importance and the majority are less important. These vertices and subgraphs have been studied extensively in online networks [37,8], but not with the idea of using them for graph synopses. Following much of this work, we choose four standard definitions of importance: degree, betweenness, closeness and PageRank. We demonstrate empirically for a number of representative online networks that these subsets of vertices do not depend highly on the choice of importance measure. Next, we show that it is possible to glean accurate information about the communication, relationship, and flow of information on the original graph and among the top vertices simply from a subgraph constructed from the important vertices. Furthermore, these properties are consistent, regardless of the importance measure we use. Finally, we analyze the evolution of the subgraph induced by the important vertices in several data sets that incorporate temporal information. We show that these synopses are persistent over time and consist of more dynamic vertices than the remaining unimportant vertices[1]

Rather than compressing the entire graph, other work has examined the utility of sampling a subset of vertices, and examining whether the resulting subgraph mirrors the original large graph in aggregate properties such as average path length, degree distribution, and clustering [19,17,4]. Sampling may be used when it is impossible to access the entire network; e.g. when one is crawling online data, or when the graph is too large to efficiently measure in its entirety. Sampling methods include node, edge, and random-walk based sampling. Other approaches include mining a subgraph for visualization of the original graph [11,36], placing sensors to detect information flow [21], constructing a synopsis by projecting queries [18], and quantifying the extent to which important vertices hold online social networks together [25].

However, all of the previous work has focused on keeping or representing the properties of the original networks; i.e. studying the entire networks. We study the more fundamental properties of the subgraphs of important vertices themselves. For example, if we identify a set of important web pages, where importance may be measured in terms on indegree or PageRank, we are not asking whether this set of web pages is a good representation of the structure of the entire web graph. Likely it is not a good representation. Rather, we are asking how those important web pages relate to one another: Do they form a connected component? How far removed are they from each other when one must navigate through other important pages? Is their relative importance preserved in the subgraph? These questions have not been studied previously because the focus has been on relating the subgraph properties to the properties of the entire original network. Instead, we consider the relationship between these important vertices to be of interest in and of itself..

Previous work has examined the evolution of large-scale networks, noting phenomena such as densification [20]. Compression has also been studied in the context of an evolving online social network [7]. However, the question posed focused on the compressibility for each individual time slice, rather than asking whether the synopsis constructed in one time slice remains stable in subsequent time slices.

We give a clear, precise definition of the algorithmic problem of *vertex-importance graph synopsis* in section 2 and discuss the computational hardness of this

[1] Part of this work originally appeared in [33]. In this work, we include the work from the original conference proceedings but also present three additional contributions. First, we investigate the time dependent properties of the graph synopses and how subgraphs of important vertices change over time using several additional data sets. Second, we examine the robustness of the difference importance measures with respect to a particular instance of an anomalous hub in one of our data sets. Third, we derive NP-hardness and NP-completeness results for the graph compression problem.

problem in section 4. We show in sections 3 and 4 that most online networks are far from the worst-case graph; they exhibit features (e.g., power-law degree distribution, short average diameter, and high clustering) that allow us to efficiently compute a graph synopsis. Moreover, we tie properties of the subgraphs to measures, such as assortativity, of the original networks. We match the empirical observations to analytical results in section 5. Finally, in section 6, we examine the time dependent properties of the subgraphs, their persistency and evolution.

2 Preliminaries

2.1 Importance measures

The definitions of *importance* or *prominence* on vertices vary significantly depending on the specific network and application. Most such measures describe the topological location of the vertices. We choose four of the most commonly used measures in various applications as our objects of study: *degree*, *betweenness*, *closeness*, *PageRank*.

Let the graph $G(V, E)$ have $|V| = n$ vertices, the four importance values defined on vertices v_i are listed below:

1. **Degree** $D(v_i)$: a measure of how many vertices in G are connected to v_i directly. If G is a undirected graph, then $D(v_i)$ is the number of undirected edges incident to v_i; if G is a directed graph, then $D(v_i)$ is the sum of indegree and outdegree of v_i, where indegree is a count of the number of directed edges to the vertex, and outdegree is the number of directed edges from that vertex to others. Degree reflects a local property of the vertices in the graph.
2. **Betweenness** $B(v_i)$: a measure of how many pairs of vertices go through v_i in order to connect through shortest paths in G:
$$B(v_i) = \sum_{j<k} g_{jk}(v_i)/g_{jk}$$
where g_{jk} is the number of shortest paths linking vertices j and k; and $g_{jk}(v_i)$ is subset of those paths that contain vertex v_i. For a directed graph G, the shortest paths are directed shortest paths. Betweenness reflects a global property of the vertices in the graph.
3. **Closeness** $C(v_i)$: a measure of the distances from all other vertices in G to vertex v_i:
$$C(v_i) = \left[\sum_{j \neq i} d(v_i, v_j)\right]^{-1}$$
where $d(v_i, v_j)$ is the distance between v_j and v_i. Intuitively, closeness means that vertices that are in the "middle" of the network are important. For a directed graph G, the closeness of a vertex could be computed in three ways: all directed paths *to* the vertex, all directed paths *from* the vertex, and all paths regardless of direction. Here we use this third version, effectively treating the graph as undirected.
4. **PageRank**: a variant of the Eigenvector centrality measure and assigns greater importance to vertices that are themselves neighbors of important vertices [29].

2.2 Description of network data sets

We chose our network data sets to be representative of web and online social network data for which one might be interested in examining the properties of important vertices and their graph synopsis. We complement the empirical data sets with analysis of Erdös-Renyi (ER) random graphs, in order to discern interesting features in real world graphs from patterns that may arise by chance. For directed and undirected graphs, we measure the properties of the directed or undirected versions respectively, restricting ourselves to the largest weakly connected component.

Erdös-Renyi random graph. An Erdös-Renyi random graph is a prototypical random graph with each pair of vertices having an equal probability p of being joined by an edge. In our model, we set the number of vertices $|V| = 10000$ and choose $p = \frac{1}{1000}$, so the average degree is $\langle d \rangle = p \times |V| = 10$.

Budyzoo data set. The first real-world network we consider is derived from the website `buddyzoo.com`. The system, no longer active, allowed users to submit their AOL Instant Messenger (AIM) buddy lists to compare with others. By treating each registered user as a node and their Buddy List as a series of edges to other nodes, a graph is formed. Our anonymized snapshot of the data from 2004 includes 140,181 registered users [14]. In this paper, we keep only reciprocal ties (74.7% of the total edges), producing an undirected graph.

TREC. The second real-world graph considered is a network of blog connections, the TREC (Text REtrieval Conference) Blog-Track 2006 data set [24]. It is a crawl of 100,649 RSS and Atom feeds collected over 11 weeks, from December 6, 2005 to February 21, 2006. In our experiments, we removed duplicate feeds and feeds without a homepage or permalinks. We also removed over 300 Technorati tags, which appear to be blogs, but are in fact automatically generated from tagged posts, and so are not true indicators of social linking. The TREC data set contains hyperlinks of various forms, including

	Erdös-Renyi	`BuddyZoo`	`TREC`	Web	`Honda-tech`
Vertices	10,000	135,131	29,690	152,171	45,718
Edges	49,935	803,200	195,940	1,686,541	459,243
ASP	4.26	5.96	3.72	3.48	3.72
Directed	False	False	True	True	False

Table 1 The average shortest path (ASP) and other characteristics of the largest components of the graphs.

blogrolls, comments, trackbacks, etc. There are 198,141 blog-to-blog hyperlinks in total, and 33,385 blogs having at least one such link.

Web graph data set. The web graph data set was collected in 1998 by Alexa[2] and has previously been analyzed as part of the "Web in a box" project at the Xerox Palo Alto Research Center [1]. Since the snapshot was collected such a long time ago, it contains *only* 50 million pages and 259,794 websites. This "small" size allows us to comprehensively analyze the web graph. We construct a directed graph where Site A has a directed edge to site B if any of the pages within A point to any page within site B.

Due to the similarity of results for the recent blog data sets and the decade old website-level data set, we expect our results to be applicable to larger, more current webcrawls.

The above datasets lack time-resolution. However, most, if not all, large-scale networks are highly dynamic, and an important question is how VIGS perform as a network grows through the addition of nodes and edges. We therefore supplement the above analyses with two time-resolved data sets:

Honda-tech. To study the evolution over time of the important vertices in a large graph, we use the Honda-tech data set which is an online forum for Honda customers to provide and exchange information and resources[3]. It has 86,000 threads and about 45,000 users from 2001 to 2008. We consider as vertices in our network construction individual users. Because users reply to posts in the forum, we have explicit relationships amongst them and we consider these as ties or edges. In order to see the dynamics of this social network, we divide the data set into 30 time snapshots. Every snapshot has a time window of approximately 3 months. Over the entire period encompassed by the crawl there are 45,718 users and 459,243 undirected edges. 99.26% of these users are situated in the largest connected component and the average shortest path length separating them is 3.72. While we performed analysis on several other evolving data sets from fora data, their results are not substantially different from those of Honda-tech, so we do not include them. We also note that this data set is studied extensively in [35] which the interested reader can consult for more information.

[2] `www.alexa.com`

[3] `www.honda-tech.com`

Second Life transaction graph. Second Life is an online virtual world with hundreds of thousands of active users. Although the environment is created by the company Linden Lab, much of the content, from buildings to landscaping to clothing to entertainment and music is created by the users themselves. The users trade objects and services using Linden, a virtual world currency that has an exchange rate with the US dollar. Such activity can be profitable, with 150M USD in user-to-user transactions taking place in the third quarter of 2009 [23]. We limit ourselves to the set of 1,490,159 users who were in the seller role at least once from July 2007 to April 2009. We also only consider transactions that were gifts or object transfers, omitting membership dues, data upload fees, classified ads., etc. This data is a specific subset of all the economic activity that takes place in SL.

3 Important vertices

In this section, we examine the graph synopsis consisting of important vertices in the network. First, we describe some properties of the entire networks. Second, we analyze the subgraphs induced by important vertices. Finally, we compare some properties of the important vertices in the subgraphs and the entire networks.

3.1 Network properties and important vertices

Degree distributions. We plot the cumulative degree distributions of three real online networks in Figure 1. We treat the Web and `TREC` networks as directed graphs and plot the distributions of their in-degrees and out-degrees. We treat `BuddyZoo` as an undirected graph. By fitting the distributions of in-degree of Web and `TREC` with power-law distributions, we get their power-law exponents, which are 2.47 and 2.16 respectively. Moreover, we can see that the degree distribution of `BuddyZoo` has a very sharp drop off at the tail, which is observed in many social networks, e.g. co-authorship networks [26]. This places blog links, a form of social linking, somewhere between navigational/informational general linking on the Web and the reciprocal, communicative linking of a social network. The distributions of out-degree of Web and `TREC` show mild deviations from power laws, consistent with other web measurements [32] and might be due to the limitation of the data sampling [34].

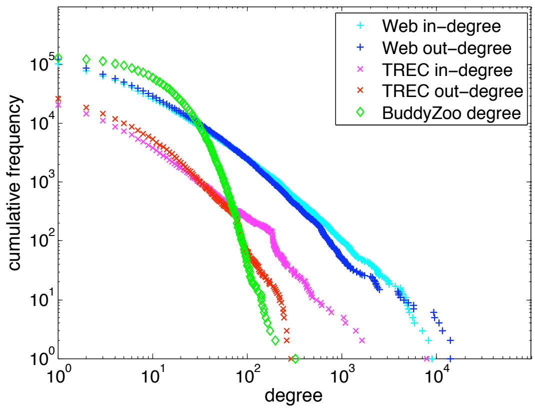


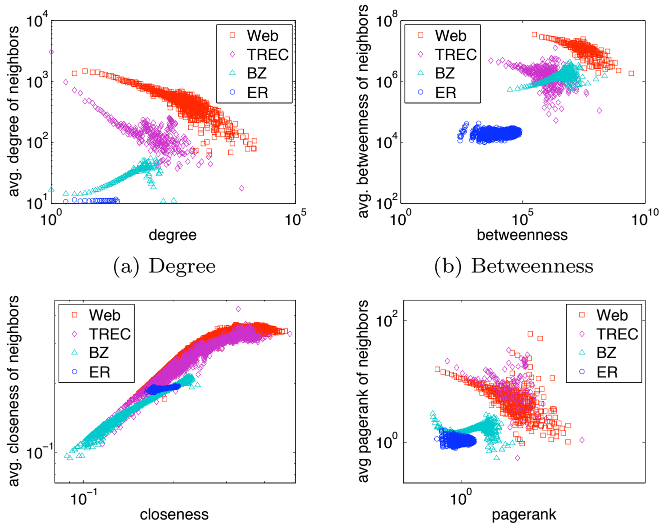
Fig. 1 The degree distributions of online networks of `BuddyZoo` data, `TREC` blog data and Web data.

Correlation of importance values of different measures. Before examining the important vertices in the networks, we look at the relationships of importance measures in different networks. Table 2 shows that all of the importance measures are positively correlated in all four networks. The two undirected graphs, Erdös-Renyi and BuddyZoo, have more highly correlated importance measures. Perhaps the directed edges of the other graphs add complexity to centrality measures. Furthermore, we see that for all of the networks, *degree*, *betweenness* and *PageRank* have higher correlation than *closeness*. Thus, we see that there are various ways of defining importance in the networks and the most *central* vertices according to different centrality measures share overlap significantly.

Correlation	Erdös-Renyi	`BuddyZoo`	`TREC`	Web
Deg, Bet	0.9920	0.8137	0.7872	0.6178
Deg, Clo	0.9474	0.7849	0.3835	0.7869
Deg, PR	0.9952	0.9486	0.7058	0.5175
Bet, Clo	0.9673	0.7541	0.3120	0.4709
Bet, PR	0.9823	0.8439	0.7439	0.6757
Clo, PR	0.9154	0.6418	0.1086	0.3253

Table 2 Spearman correlations between importance measures of vertices. All the p-values of the correlations are < 0.0001.

Assortativity. The concept of *assortativity* or assortative mixing is defined as the preference of the vertices in a network to have edges with others that are similar. Here, we will focus on similarity with regard to centrality. We choose to measure the average value $\langle k \rangle$ of the neighbors of vertices of importance value k, i.e. $\langle k \rangle_{neigh}(k) = \sum_{k'} k' P(k'|k)$, where k is determined by each of the four different importance measures [31]. From the change of $\langle k \rangle_{\mathrm{neigh}}(k)$ as k increases, we deduce the network's assortativity for this particular valuation. When the overall slope of $\langle k \rangle_{neigh}(k)$ is positive, the network is assortative; if the overall slope is negative, then it is disassortative. Otherwise, the network is neutral (e.g. the assortativity of degree of Erdös-Renyi random graphs).



(a) Degree

(b) Betweenness

(c) Closeness

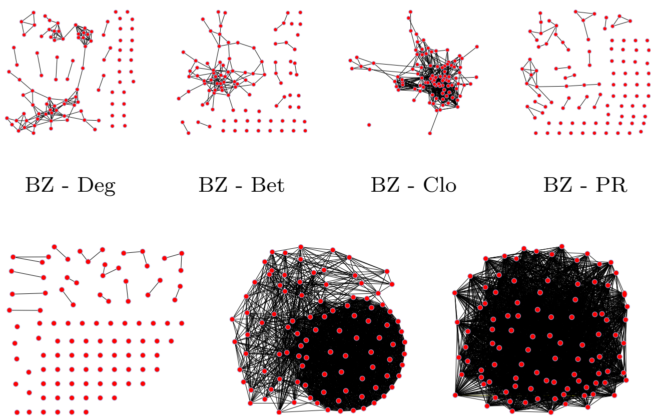
(d) PageRank

Fig. 2 The slopes of the distributions of $\langle k \rangle_{\mathrm{neigh}}$ show the assortativities.

In Figure 2, we can see that all four networks are consistently assortative with regard to the closeness importance measure. This confirms our intuition. Because closeness reflects the average distance of a vertex to all others in the graph, a neighbor of a vertex with high closeness is at most one additional step removed, and hence must also have high closeness. The other three importance measures consistently show that the Erdös-Renyi random graph is a neutral graph, that `BuddyZoo`, similar to other social networks[27], is assortative, and that the Web and `TREC` blog networks are mildly disassortative. We'll see in Section 5 that this result does not mean that important blogs avoid linking to other important blogs. Rather, there is such a large skew in the degree distribution, that in order for a high degree vertex to link to such a large number of others, it must form edges to many of the more abundant low degree vertices.

3.2 Important vertices in their subgraphs

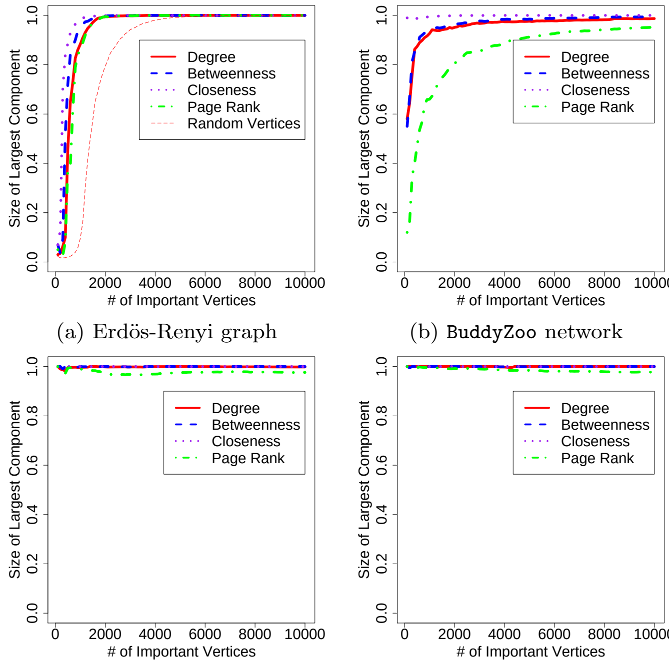
In this section, we discuss important vertices and the subgraphs induced by these vertices. Such analysis helps us to discover the information hidden behind the important vertices in the real online networks, and how we can utilize them for graph synopsis. We do not fix a specific threshold for inclusion of important vertices in the subgraph, as this may vary by application. Rather, in our study what occurs as we allow the absolute number of important vertices, m, to vary, as long as $m << n$, where n is the number of vertices in the original network.



BZ - Deg BZ - Bet BZ - Clo BZ - PR

ER - Deg `TREC` - Deg Web - Deg

Fig. 3 In the top row are subgraphs induced by the top 100 important vertices of `BuddyZoo` for all four importance measures, while in the bottom row are subgraphs induced by the 100 highest degree vertices in the other three networks.



(a) Erdös-Renyi graph (b) `BuddyZoo` network (c) `TREC` blog network (d) Website network

Fig. 4 The sizes of largest connected component of the subnetworks of important vertices in an Erdös-Renyi random graph and three real online networks.

Figure 3 shows the subgraphs induced by the four importance measures in `BuddyZoo` and the highest degree vertices in the other three networks. We observe that these subgraphs can be markedly different for different measures of importance, even within the same graph, in spite of high correlation in importance measures among vertices. They may also vary significantly between graphs, even for the same importance measure. There are several explanations of this behavior. Given the high assortativity of the closeness measure, we are unsurprised to find that individuals of high closeness are closely connected in the BuddyZoo graph. Buddyzoo also has individuals of high degree, but there were limits imposed on the number of contacts one has both by AOL and individuals' own bandwidth, and so the largest connected component among high degree vertices does not contain all such vertices. On the other hand, the highest degree vertices in both the blog and web data sets have such high degree that they tend to form a single connected component.

Connectivity. The first question we address is whether the connectivity of important vertices depends on other, less important, vertices or whether they are already well connected through one another. In the Erdös-Renyi random graph, the size of the largest connected component is given by the solution x to the equation

$$x = 1 - e^{-\langle k \rangle x}$$

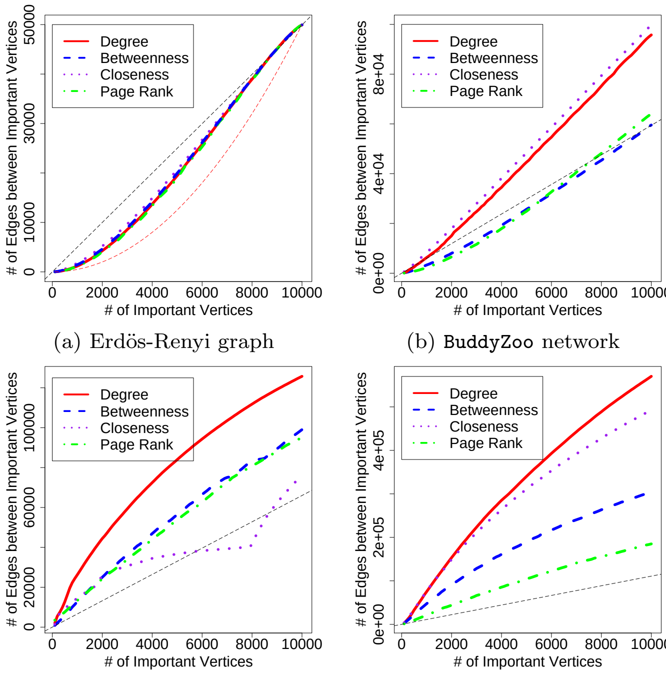
where $\langle k \rangle$ is the average degree of the graph. The solution to this equation, shown as a dotted red line in Figure 4(a), represents the change in size of the largest connected component of the subgraphs induced by picking vertices randomly from the Erdös-Renyi random graph. When we choose vertices according to importance instead, the subgraphs have significantly better connectivities, with the largest connected component occupying 96.5% of the subgraph once the subgraph contains over 15% of all vertices in the graph (i.e., 1,500 important vertices vs. 10,000 total vertices).

Moreover, from Figure 4, we see that the important vertices are even more highly connected in the real networks. For `BuddyZoo`, more than 95% of the important vertices of highest degree, betweenness or closeness are in the largest connected component when they comprise just 1% percent of all vertices in the network (i.e., 1,500 important vertices vs. 135,131 total vertices). In addition, more than 95% of the 10,000 highest PageRank vertices are in the largest connected component. For both of the two directed networks, the `TREC` blog network and the network of websites, the most important vertices are very well connected (> 99.5%) even when their numbers are very small (< 0.05% of all the vertices in the networks). Note that this very high level of connectivity is in spite of the dissortative nature of the TREC and website networks with respect to degree, betweenness and PageRank, where important vertices tend to connect to less central vertices. We can reconcile the two by observing that the important vertices are already interconnected, so the negative assortativity comes from highly connected vertices being connected to lower degree vertices simply because they already have so many connections and there is only a small percentage of vertices of similarly high degree [30].

Density. The previous observations tell us that the connectedness of important vertices is high even when we omit all other vertices in the original graph and even when they comprise a very small fraction of the entire network. Next, we examine just how dense their connections are. In Figure 5, we show the relationships between the number of edges incident on important vertices and the number of important vertices.

Figure 5 (a) shows that for an Erdös-Renyi graph, the important vertices according to all four measures have a higher average degree in the subgraph than randomly chosen vertices (red dashed line), but this average degree is lower than the average degree in the complete graph (black dashed line). The average degree of the graph reaches a maximum when all of the vertices in the graph are included. Moreover, from the direction of the curves, we can see that the number of edges e increases super-linearly with the number of important vertices n, i.e. $\Theta(n) < e < \Theta(n^2)$.

However, Figures 5(c) and (d) reveal the opposite behavior for networks with highly skewed degree distributions (TREC and Web). The curves of each network do not overlap as much, and the average degree of the important vertices in the subgraph is higher than the average degree in the original network. This indicates that rather than being sparser, as was the case for the Erdös-Renyi subgraphs, the subgraphs of important vertices in real world online networks are actually *denser*. Finally, for the BuddyZoo network (Figure 5 (b)), which is assortative, but not power-law in degree, we see a mix of trends. Subgraphs of vertices



(a) Erdös-Renyi graph (b) `BuddyZoo` network (c) `TREC` blog network (d) Website network

Fig. 5 The growth of numbers of edges between important vertices. The slope of the black dash line in each plot is the ratio of the number of edges vs. the number of vertices in the entire network.

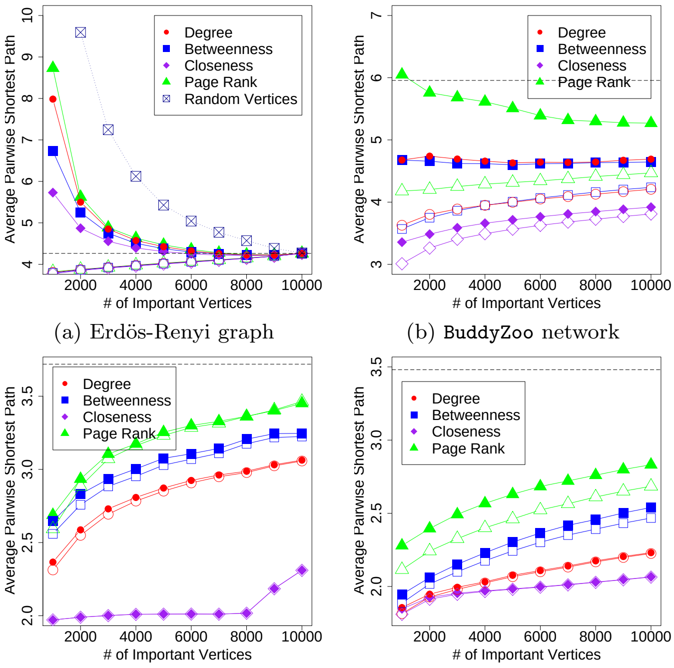
with high betweenness and PageRank tend to be a bit sparser than the complete network, but the most important vertices according to degree and closeness are more densely connected (this is also apparent in the visualizations in Figure 3).

In examining these real online networks, we see that although the densities of connection among important vertices vary considerably in different networks with different importance measures, in general, they are significantly denser than for subgraphs of randomly chosen vertices in the Erdös-Renyi random graph.

3.3 Original vs. subgraph properties

Distance. In Section 3.2 we saw that even without any additional vertices from the original graph, the subgraphs of important vertices in the three online networks are already well connected. Next we examine the second property that we want to preserve for our graph synopsis problem: the average shortest paths (ASP) between reachable pairs of important vertices.

Figure 6 shows the comparison curves of ASPs of important vertices in their induced subgraphs and in the original networks. In the Erdös-Renyi random graph, the ASP between important vertices utilizing the entire graph is on average shorter than the ASP between



(a) Erdös-Renyi graph (b) BuddyZoo network

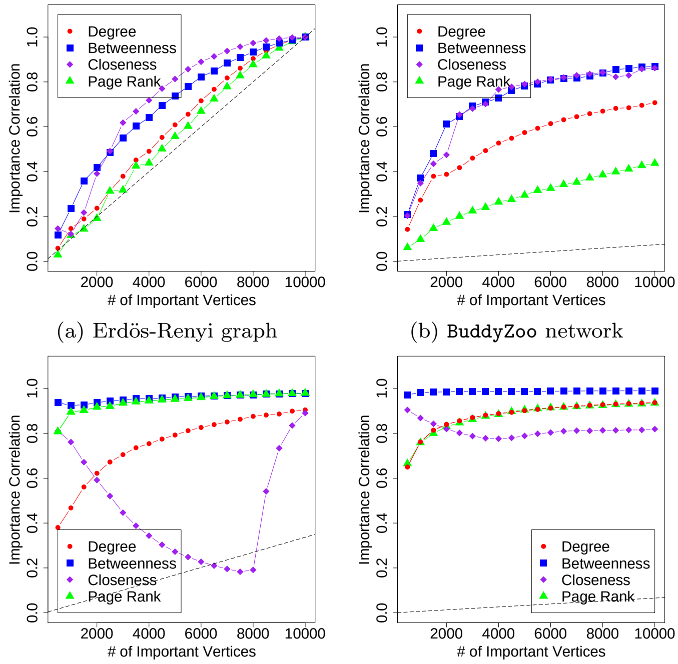
(c) TREC blog network (d) Website network

Fig. 6 The ASP of: all vertices in the entire networks (black dashed line), important vertices in the the subgraphs (solid points), important vertices in the entire networks (hollow points).

all pairs of vertices (the dotted baseline). But when the ASP uses only routes in the induced subgraph there are significantly more hops between important vertices on average, which indicates that in random networks they are not closely connected, and their shortest paths route through non-important vertices. Nevertheless, subgraphs of important vertices in ER graphs have slightly shorter ASPs than subgraphs of randomly selected vertices.

In contrast to the Erdös-Renyi random graph, all three real online networks consistently show that the ASPs between important vertices are much shorter than the average shortest paths of the entire network; and almost all of them are increasing as more vertices are added in by lowering the importance threshold. What is more, by comparing the ASPs of important vertices in the original graphs and in the subgraphs, we see that their values are extremely close in most cases, especially for the `TREC` and Web data, e.g. the solid and hollow purple points (ASP of vertices of highest closeness) are almost exactly overlapped. This indicates that important blogs are most efficiently connected through other important blogs.

Relative importance. In addition to the connectedness of important vertices, we are also interested in their relative ranking: if we only keep the important vertices and the edges among them, how would the vertices rank in the new subgraph with the same importance measure? In order to answer this question, we generate subgraphs of different sizes for all networks. We then compute the importance of the vertices in the subgraphs according to the same importance measure used to select them. Finally, we compute the Pearson correlation of the importance values of those vertices in the original graph and in the subgraph.



(a) Erdös-Renyi graph (b) BuddyZoo network

(c) TREC blog network (d) Website network

Fig. 7 Pearson correlations of importance values of vertices in subgraphs and original graphs. The black dashed lines are the base lines starting from 0 when the number of vertices is 0; and ending at 1 when all the vertices in the networks are included.

Figure 7 shows that the correlations are all much higher for the real-world online networks than the Erdös-Renyi random graph, and that this is especially true for the Web and TREC data. The high correlations of the online networks tell us that the ranking of importance in the subgraphs of important vertices is highly consistent with their ranking in the original graphs. This suggests that, e.g. it may not be necessary to crawl all blogs to get an accurate ranking of the most important blogs. Rather, the links among the top blogs themselves may already provide fairly close approximate rankings.

There is a clear anomaly, however, in the correlations of importance values for closeness for the TREC dataset, as visible in figure 7(c). Figure 5 also exhibits an anomaly in the density of closeness-induced subgraphs. There is an odd dip or jump in the figures near the 8000 vertices mark. A closer examination of the TREC graph reveals the culprit: a single vertex with connections to 99% of the first 7961 high-closeness vertices. We found that this particular vertex was a blog serving as an aggregator of content on other blogs. Its indiscriminate linking caused it to connect not just to blogs worthy of linking, but also to many new, private, and spam blogs. A large number of the vertices connected to this high-degree vertex obtain high closeness through this connection, since they are at most distance 2 from any other node linked to the aggregator. However, they do not necessarily have many other connections, which is reflected in the graph synopsis as well.

To understand this behavior we examined the number of edges in the subgraph as the closeness importance threshold is lowered and more vertices are added to the synopsis. We replot the TREC curve from Figure 5 in Figure 8 and highlight 4 different regions corresponding to 4 groups of nodes that are successively incorporated. Nodes in group 1 between about 1 and 2000 are connected to the aggregator and to many other high-closeness nodes as well. Therefore as we include more of them in the subgraph, we are also capturing more of a dense, central region in the original network. These connections lessen in group 2 until we reach group 3 in the 6000-7000 range, nodes with only one connection (to the aggregator) are being added. Suddenly, nodes with no connection to the aggregator are added again, and the number of edges between important vertices resumes its rapid increase in group 4, since now closeness once again reflects a higher degree of connectivity in general, and not just to the central aggregator.

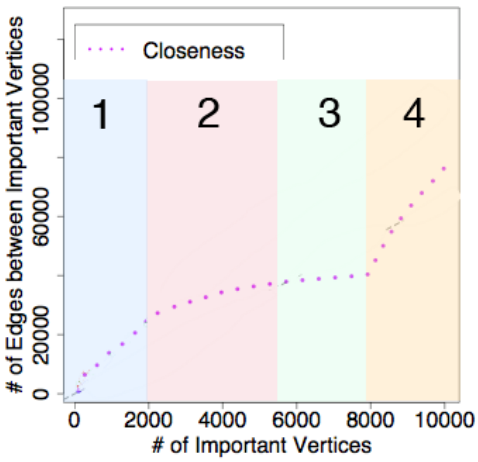


Fig. 8 The number of edges between important vertices as a function of the number of important vertices included in the subgraph for the TREC data set and the closeness measure only. The different regions in number of important vertices are highlighted.

The aggregator's presence is notable in other graph properties as well. In Figure 6(c) we see the average shortest path as a function of the number of important vertices in the synopsis (for four different importance measures). Observe that for the closeness measure, the average shortest path stabilizes at 2, as most nodes beyond the first few thousand have no connections to other important vertices, but may reach any directly through the aggregator. The ASP length begins increasing as nodes not connected to the aggregator are added.

The correlation between importance in the subgraph versus in the full graph is also affected. Blogs connected to the aggregator have their closeness score inflated, until better-connected blogs are added past the 7961 mark. The aggregator serves as a bridge for its connections, giving them a very short (2 hops) path to many other nodes. This inflates their closeness score dramatically. On the other hand, it does nothing for the degree score, as this is only one link. A node's gain in PageRank from its connection to the aggregator is diluted by the aggregator's high degree. While the aggregator itself will have a high betweenness score, it confers none of this to its neighbors, as any node may be used to pass through the aggregator.

The above discussion of the effect of a single aggregator vertex on VIGS using a specific importance measure illustrates how vertex importance synopses can call attention to key features of the network, as well as a demonstration of the benefit of using multiple measures of importance to capture a robust and meaningful subgraph.

3.4 Summary

After studying the important vertices and their induced subgraphs, we can make two overall observations about the four networks: (i) Different importance measures yield subgraphs of varying density and topology as is evident in Figure 3. (ii) However, in spite of these differences, "important vertices" in the online networks have some properties that agree with each other, which are essential for the graph synopsis we are looking at: they connect to each other more directly than average; their distances to each other are closer than random vertices; and their relative ranks are positively correlated to their importance ranks in the original networks. Thus, we know that in the real online networks, in contrast to random graph model, the subgraphs induced by the important vertices tend to preserve information about the relationships among important vertices, and we can use the subgraphs to study the properties of important vertices in the original graphs.

4 Compression with guarantees

While retaining only the important vertices may be sufficient to capture most of the relationships among them in real-world networks, in general we have no guarantee that these induced subgraphs preserve any properties at all (whether of the important vertices or of the original graph). We cannot even guarantee the most basic property of connectivity of the important vertices. In this section, we rigorously define the graph compression problem, analyze the computational complexity of two heuristic algorithms, and discuss the trade-offs of these approaches.

4.1 Hardness of compression with guarantees

We define the Basic Graph Compression Problem as follows: In a connected unweighted graph $G(V, E)$, every vertex is assigned an importance value. Taking the original graph $G(V, E)$ and the set of vertices S with largest importance values as inputs, find the minimal set of additional vertices ν, which form a connected subgraph $G'(V', E')$, where $V' = S + \nu$ and $V' \subseteq V$, $E' \subseteq E$.

We recall the Network Steiner Tree Problem which is NP-complete [10]. Let $G = (V, E, d)$ be a graph with nonnegative edge lengths $d : E \to R^+$ and let S be a set of distinguished vertices. A Steiner tree is a tree which spans all members of S, and possible with additional vertices in V. The problem asks for a minimum cost Steiner tree T_{min}, where the cost of a set of edges is the sum of lengths of its elements. A heuristic method, the*Minimal Spanning Tree* algorithm gives solutions to this problem with approximation ratio 2 [12].

Theorem 1 Basic Graph Compression *is NP-hard.*

Proof We prove that Basic Graph Compression is NP-hard by proving that Network Steiner Tree Problem is polynomial-time Turing-reducible to it.

Assume that $\eta(g = (\tilde{V}, \tilde{E}, \tilde{d}), \tilde{s})$ is an instance of the Network Steiner Problem. First, we transform the domain of $\tilde{d}$ from nonnegative real numbers to nonnegative integers by multiplying them by a fixed constant value so that they are all in the new domain $\tilde{d'} : \tilde{E} \to Z^*$.

Then we convert the weighted graph g into a unweighted graph g' by replacing every edge $e_\omega \in \tilde{E}$ whose weight is $\omega \in Z^*$ with a chain of $\omega - 1$ vertices and ω unweighted edges. The first and last edges on this chain are connected with the vertices initially incident on e_ω. After this transformation, we get an unweighted graph $g'(\tilde{V'}, \tilde{E'})$, where $|\tilde{V'}| = |\tilde{V}| + \sum_{|\tilde{E}|}(\omega - 1)$ and $|\tilde{E'}| = \sum_{|\tilde{E}|} \omega$.

Now we have a unweighted graph $g' = (\tilde{V}', \tilde{E}')$, and we will prove that the set of vertices $\nu \subseteq \tilde{V}$ is the solution of η if and only if the corresponding set of vertices after transformation is the solution of the graph compression problem $\tau(g' = (\tilde{V}', \tilde{E}'), \tilde{s})$.

Assume that ν_0 is the set of additional vertices which is the solution for η; i.e. the Steiner tree of instance η is the minimum spanning tree with the set of vertices $s + \nu_0$, and the set of edges E_0 whose sum of weights is w_0. After transforming g to the unweighted graph g', this Steiner tree is transformed to another tree t' with the set of vertices $s+\nu_0+\alpha$, where $|\alpha| = w_0 - |E_0|$. Thus, the number of total vertices in tree t' is $|s + \nu_0 + \alpha| = |s + \nu_0| + w_0 - |E_0|$. Moreover, it is easy to see that $|s + \nu_0| = |E_0| + 1$, so $|s + \nu_0 + \alpha| = w_0 - 1$. Since we know that w_0 is the weight of the Steiner tree of η, we know the minimum number of additional vertices added to the tree t'. Thus, this is the solution to our graph compression problem τ.

It is similar to prove that if a tree is not the Steiner tree of the problem η, then it is also not the solution tree to the corresponding graph compression problem.

Thus, we have proved that the the basic version of graph compression problem is NP-hard.

Corollary 1 Basic Graph Compression *is NP-complete.*

Proof Using a similar proof to that above, it is easy to prove that the basic version of graph compression problem is polynomially time Turing-reducible to the Steiner Tree Problem, which is NP-complete.

4.2 Heuristic algorithms

There are, however, several heuristic algorithms that guarantee the preservation of some properties of the important vertices in the original graph. We detail the KeepOne and the KeepAll algorithms [11] next, and note the similar web projection method [18].

KeepOne. Let K_1 be the set of important vertices, the goal is to find the minimal set K_2 such that there is a tree induced by $K_1 \cup K_2$. The approximation algorithm is first to build a minimum spanning tree on the complete graph on K_1 where an edge (u, v) has weight equal to the length of a shortest path from u to v. The set K_2 consists of any additional vertices along any "path" edge in the minimum spanning tree. The result is the graph induced by the vertices $K_1 \cup K_2$.

The KeepOne algorithm guarantees the connectivity of the compressed graph, has the same set of additional vertices as the projection method in [18], and only introduces more edges, which means it may have better diameter preservation than the projection method.

Unfortunately, retaining only connectivity may provide a distorted view of the original graph. We see in Figure 9 an example of a graph on n vertices in which the distance of the original vertices a and b is 3 but in the compressed graph built by KeepOne, their distance is $n-3$. The ratio of the distances is $\frac{n-3}{3}$ which we can make arbitrarily large by increasing the number of vertices n. That is, KeepOne retains connectivity but may drastically distort the distance between some pairs of important vertices. To ameliorate this problem, one can use the KeepAll algorithm [11] which keeps vertices that lie along a shortest path between any two vertices in K_1.

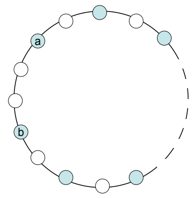


Fig. 9 The distance of important vertices a and b in the original graph is 3 and $n-3$ in the compressed graph obtained by KeepOne. The ratio of distances can be made arbitrarily large as $\lim_{n\to\infty} \frac{n-3}{3} = \infty$.

4.3 Empirical evaluation and trade-offs

While Figure 9 shows that the worst case distance preservation of KeepOne may be arbitrarily bad, real-world networks are far from the worst case. Furthermore, the KeepOne and KeepAll algorithms illustrate that there are some tradeoffs we may make in compressing real-world graphs—we can maintain distances at the cost of keeping a few additional vertices. To explore these empirical tradeoffs, we apply both the KeepOne and KeepAll algorithms to three networks. Table 3 shows these results. Since the results with the Web data are very similar to `TREC`, we do not list them here for conciseness. From the table, we can see that if we insist on preserving the pairwise shortest paths of all important vertices, we must include many more additional vertices (thus increasing the size of our synopsis). Furthermore, we must do so even though the average pairwise shortest paths in the subgraph of just the important vertices is already close to that of the original graph. Note that we increase the size of the synopsis by fewer than 100 additional vertices when we preserve connectivity (with KeepOne), but we need over 3000 additional vertices when we also insist on preserving distances. In short, while the problem of preserving connectivity in graph compression is NP-complete, heuristic algorithms such as KeepOne can preserve connectivity with a lower cost, while preserving the distances demands quite more. In this sense, we can also see that the short pairwise shortest paths of important vertices in their subgraphs and their original graphs is a special and important property of the online networks we study.

5 Analytical Discussions

In this section we present the expected density of subgraphs of random graphs with varying degree distributions, in order to contrast these expected values with the empirically observed measurements. We limit ourselves to vertex degree as the sole importance measure and assume that the graphs are random aside from the degree distribution, which we specify. We then obtain the density of the subgraph by deriving the probability that an edge in the original graph lies between two vertices in the subgraph.

First, we find the degree k_i of the least important vertex among the set of top i most important vertices. We do so by calculating the expected number of vertices of degree at least k_i in a network with $n = |V|$ vertices. Furthermore, we assume that the expected number is actually equal to i so that

$$i = n \cdot P(k_i).$$

Because we are given the ccdf $P(k)$ explicitly for Erdös-Renyi and power law random graphs, we can solve the previous equation for k_i and, after doing so, we find the probability that an edge is incident to a single important vertex, $e \to V_i$, given by

$$\mathbb{P}(e \to V_i) = \frac{1}{|E|} \int_{k_i}^{n} k \cdot p(k)\, dk$$

where $p(k)$ is the pdf of the degree distribution. Using independence of the edges, we find that the number of edges within the subgraph of important vertices is simply

$$|E_i| = |E| \cdot \mathbb{P}(e \to V_i)^2.$$

5.1 Erdös-Renyi graphs

In an Erdös-Renyi random graph, the degrees are distributed according a Poisson distribution where the probability of a vertex having degree larger than the mean decreases exponentially. As a result, even when selecting the highest degree nodes, their degree will be within an order of magnitude of the average degree $z = \langle k \rangle$ of the network.

Subgraph	Add vts	LC	Avg PSP	Subgraph	Add vts	LC	Avg PSP	Subgraph	Add vts	LC	Avg PSP
Erdös-Renyi				`BuddyZoo`				`TREC`			
Sub-Deg100	0	NA	NA	Sub-Deg100	0	0.58	NA	Sub-Deg100	0	1	1.636
KO-Deg100	80	1	14.526	KO-Deg100	33	1	9.440	KO-Deg100	0	1	1.636
KA-Deg100	3222	1	3.649	KA-Deg100	2199	1	3.233	KA-Deg100	34	1	1.609
Sub-Bet100	0	NA	NA	Sub-Bet100	0	0.55	NA	Sub-Bet100	0	1	2.085
KO-Bet100	68	1	15.497	KO-Bet100	35	1	16.087	KO-Bet100	0	1	2.085
100	3185	1	3.633	KA-Bet100	2376	1	3.171	KA-Bet100	216	1	1.994
Sub-Clo100	0	NA	NA	Sub-Clo100	0	0.99	2.599	Sub-Clo100	0	1	1.716
KO-Clo100	62	1	11.474	KO-Clo100	1	1	2.624	KO-Clo100	0	1	1.716
KA-Clo100	3000	1	3.604	KA-Clo100	531	1	2.324	KA-Clo100	0	1	1.716
Sub-PR100	0	NA	NA	Sub-PR100	0	0.12	NA	Sub-PR100	0	1	1.298
KO-PR100	87	1	15.404	KO-PR100	75	1	11.517	KO-PR100	0	1	1.298
KA-PR100	3338	1	3.672	KA-PR100	3978	1	3.880	KA-PR100	36	1	1.294

Table 3 Comparison of the properties of subgraphs generated by different methods with important vertices in Erdös-Renyi random graph, `BuddyZoo` and `TREC`. Sub-ImportanceMeasure100 is the subgraph induced by top 100 important vertices only; KO- is the subgraph generated by KeepOne; KA- is the subgraph generated by KeepAll. LC is the fraction of important vertices in the large component of the subgraph. Avg PSP is the average pairwise shortest path length in the subgraph.

In Figure 10, we show the number of edges in the subgraph of an Erdös-Renyi graph, using the normal distribution with mean z and standard deviation $\sigma = \sqrt{z/n * (1 - z/n)}$, is

$$i = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{k_i - z}{\sigma\sqrt{2}}\right)\right).$$

We see that when the number of important vertices is small, the degree within the subgraph is lower than the degree of the original graph. Using well known properties of Erdös-Renyi graphs, we expect that when the average subgraph degree is 1, a giant component will emerge in the subgraph, and further, when the average degree is $\log(n)$, the subgraph will be path connected. This is consistent with the set of connectivity and density measurements on simulated Erdös-Renyi graphs in Section 3.2.

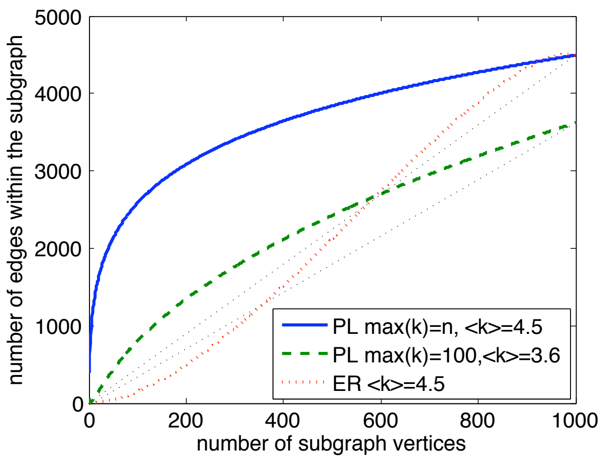


Fig. 10 The number of edges between important vertices, where importance is measured by degree, in three networks: 1) power law network with $\alpha = 2.2$, $n = 1000$, 2) Erdös-Renyi graph with the same average degree, and 3) power-law graph with the same exponent but a cutoff at $k = 100$. Two dotted lines show what the number of edges would be if the average degree in the subgraph were equal to the average degree in the original network.

5.2 Power law graphs

We expect different behavior in power law graphs, where high degree vertices are so well connected, that they will naturally connect not only to a large portion of the network, but to one another as well. For example, in a power-law graph with exponent α and no cutoff on the degree[4], one vertex on average is expected to have degree $N^{1/(\alpha-1)}$ [28]. For $\alpha = 2$, this means that one expects one node to be connected to majority of the other nodes.

In selecting high degree nodes in a power law graph, we are selecting nodes that are likely to be connected to each other by virtue of the fact that so many edges are incident on them. The number of vertices with degree

[4] a cutoff may be imposed such that $P(k) \sim k^{-\alpha}$ for $k < \max(k)$ and 0 otherwise

k_i or greater is given by

$$i = n\ P(k \geq k_i) = \frac{n}{k_i^{\alpha-1}}$$

Solving for k_i, we have that the degree of the i^{th} most important vertex is $k_i = (\frac{n}{i})^{\frac{1}{\alpha-1}}$. Next, we want to find out what proportion of the edges are incident on the i most important vertices. For this we have

$$P_e(i) = P(e \in e_i) = \frac{\int_{k_i}^{n} kp(k)dk}{\int_{1}^{n} kp(k)dk} \quad (1)$$

$$= \frac{k_i^{2-\alpha} - n^{2-\alpha}}{1 - n^{2-\alpha}} = \frac{(\frac{n}{i})^{\frac{2-\alpha}{\alpha-1}} - n^{2-\alpha}}{1 - n^{2-\alpha}} \quad (2)$$

Figure 10 shows that the average degree in the subgraphs of important vertices is actually *higher* than in the original graph. We repeat the analysis using a degree distribution cutoff $\max(k)$ that is lower than the total number of nodes n. This cutoff not only disallows very high degree vertices, but also lowers the average degree in the original subgraph. When the cutoff is introduced, the subgraph still maintains a higher average degree than the original graph, but the difference is less pronounced.

Note the similarity with Figure 5, showing the number of edges in the subgraph for the TREC and Web data sets, both of which are power law in nature (although directed). In both the analytical and empirical subgraphs, the average degree is higher than it is for the entire graph. We should mention that for exponents $\alpha \sim 2$ and very small i, Equation 2 would yield a higher average degree than there are important vertices to connect to. This is in fact a known property of random power law graphs, where simply fixing the degree of a vertex and allowing it to satisfy this degree by forming edges at random would create a non-vanishing frequency of multiple edges between highly connected vertices. If one disallows multiple edges, the networks become mildly disassortative, consistent with our empirical observations.

6 Time dependence

In this section, we analyze how stable and how active vertex important synopsis are in an evolving social network. We first analyze VIGS in the evolving Honda-tech online forum and then perform the same set of analyses for the Second Life data set. In the Honda-tech data set, each user is a vertex. If there is ever a reply from user i to user j, then there is an edge between vertices i and j. We consider the reply relationship to be symmetric in the sense that it shows that two users share some common interests or conflicting opinions [13]. The social networks constructed from this data set are undirected, simple graphs (i.e. both self edges and multiple edges are removed).

When a user first posts, we count that as the initial appearance of a new vertex. A vertex is active (after its initial period) if edges are added to it in the subsequent periods; i.e., if the user replies or is replied to by another user during those time periods. A vertex may be inactive for some time after its initial appearance before additional edges are added. Note that with this definition of activity, vertices and edges only appear, they never disappear. Vertices can become inactive with no edge updates. A vertex can still gain new edges even if the user stops posting to the forum, if her posts keep getting new replies from other users. We segment the data into 30 time periods over the eight years of data, thus every time period spans approximately 2.8 months.

We observe first that the number of vertices in the graph grows exponentially and that before the tenth time period, there are not many vertices from which to build an important subset; only after the tenth period are there over 1000 vertices. See Figure 11 for the growth in time of the graph as a whole. (See also [35] for more information.) For most of the analysis below, we begin after the fifteenth time period in order to have a large enough graph from which to extract an important subgraph.

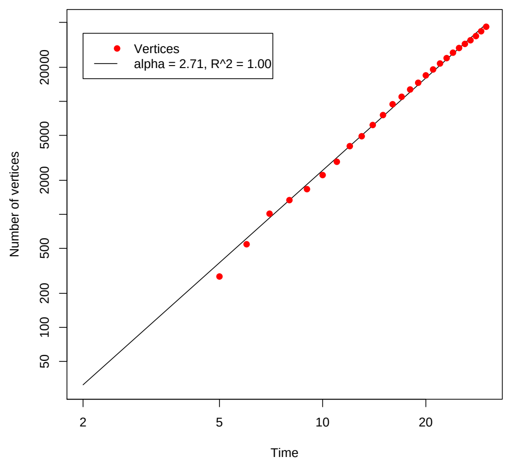


Fig. 11 The number of vertices in Honda-tech grows exponentially over time. Only after 10 time periods do we have enough vertices (over 1,000) from which we can extract an important subset.

In the Second Life data set, each user is a vertex and there is an edge between vertices i and j if a transaction takes place between the two users. We consider the transaction relationship to be symmetric and omit the transaction amount for this analysis. Like the Honda-tech social network, the social network constructed from this data set is simple, undirected, and unweighted. We use the same definition of active vertex.

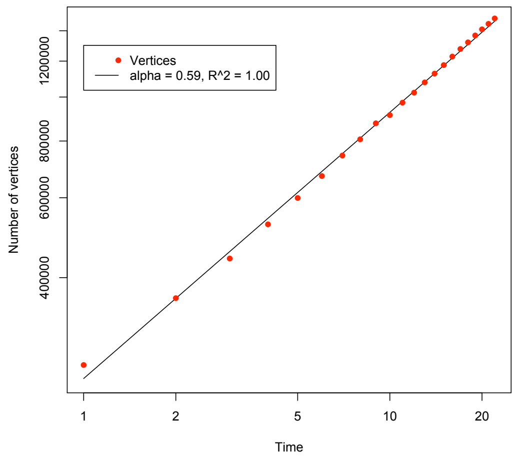


Fig. 12 The number of vertices in SL grows exponentially over time. While this growth is slower than that of the Honda-tech data set, there are considerably more vertices initially.

6.1 Persistency of important vertices

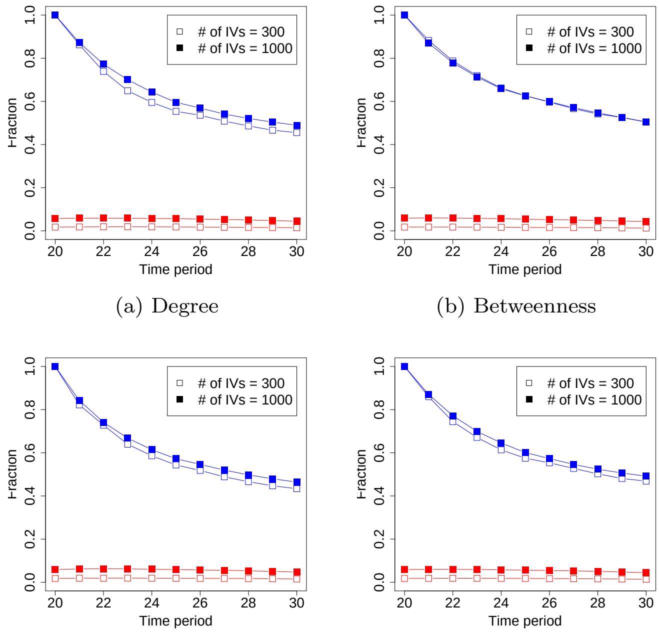
The previous analysis suggests that in a static graph important vertices are better connected than the other vertices and that their induced subgraphs preserve properties of the important vertices themselves. We do not know if these properties are persistent over time, or whether these features are simply an artifact of examining a large social network at a single point in time. To examine the persistency of the important vertices in the Honda forum data set over time and how this compares with the unimportant vertices, we select $N = 300$ and 1000 vertices with the highest importance values at time i and denote this set IV_i (we begin with the 20th time period). By comparing the set of important vertices at time i and the set of all vertices that are important up to time i, we get the blue persistency curves shown in Figure 13:

$$\frac{|\mathrm{IV}_i|}{|\bigcup_{k=1}^{i} \mathrm{IV}_i|}$$

Contrast these blue curves with the red curves that show the fraction of all important vertices up to time i in the set of all vertices that are added to the network up to time i:

$$\frac{|\bigcup_{k=1}^{i} \mathrm{IV}_i|}{|\bigcup_{k=1}^{i} \mathrm{V}_i|}$$

The curves in the figure show that while the set of important vertices stabilizes over time, in part because the entire history of a vertex is used to measure its cumulative importance at time i, these important vertices still constitute a very small fraction of the entire graph. We further varied the importance measure used to select important vertices, and the results show that the persistence of important vertices displays similar behavior across the four different measures. THe results for the Second Life data set are similar.



(a) Degree (b) Betweenness (c) Closeness (d) Page Rank

Fig. 13 The blue persistency curves show the fraction of the set of important vertices at time i in the set of all vertices that are important up to time i. The red persistency curves show the fraction of all important vertices up to time i in the set of all vertices that are added to the network up to time i.

6.2 Evolution of important vertices

Because vertices do not disappear from the graph, we measure their lifespan by counting the number of time periods after their initial appearance where they continue to receive edge updates. When we choose the top 10% vertices with highest degrees as important vertices and begin measuring the importance values at time step

15, we find that the average staying time for important vertices is 10.661 time periods and the average staying time for unimportant vertices is 8.950 time periods. The difference is significant ($p-$value < 0.05); important vertices have a longer active time spans than unimportant ones. Similar results hold for other importance measures. We also notice the U-shape distributions of the staying time lengths of both important and unimportant vertices, shown in Figure 14. A vertex is likely to remain active for just an additional 0, 1 or 2 time periods because most of the users never come back after being active for short time. On the other hand, the high probability at several of the longest staying time lengths is due to the limitation of the time window of the data set. Vertices that are observed as active at the last time step, or next to last time step, are likely to have remained active after the observation period. We believe the U-shape is more reflective of the users' posts receiving subsequent replies, rather than a user themselves either posting only during a very brief time period, or throughout. Because many of the threads on the Honda forum remain active for years, as users discuss car troubles for cars of a given vintage, a vertex may stay "alive" for years, even without posting, if their early posts prompted long lasting discussion.

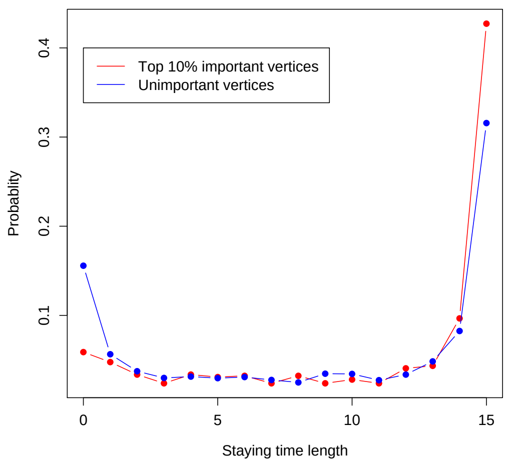


Fig. 14 The distributions of staying time length of important vertices versus unimportant vertices at time step 15. The important vertices are the top 10% vertices with highest degree at time step 15.

A second evolutionary feature of important vertices we examine is the rate of edge updates after the vertex' initial appearance—once a vertex appears, how frequently do we see edge updates for important versus unimportant vertices and how long after their initial appearance do we see many edge updates. Figure 16

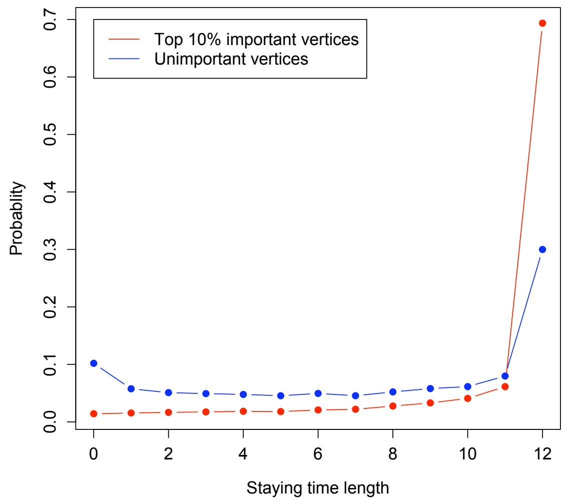


Fig. 15 The distributions of staying time length of important vertices for Second Life, also at time step 15. The important vertices are the top 10% with highest degree at this time.

shows the frequency of edge updates for (degree) important and unimportant vertices. The importance is measured at the last time step and we take the top 10% vertices with highest degrees as important vertices. We note that the important vertices are more likely to have continuous edge updates than unimportant vertices. Thus, not only are the important vertices more stable, and active for longer time periods, but they acquire edges at a higher rate. This implies that by observing important vertices over time, we are capturing more of the activity in the graph than if we were to pick just a random subset of vertices to follow.

7 Related Work

In this section, we examine the graph sampling problem and the rich-club phenomenon. Both of them have some similarities with our problem: the former also studies how to get "good" subgraphs given large massive networks; and the later focuses on the set of "important vertices". However, they are still different from our problem in various aspects. In graph sampling, one aims to devise a sampling method, e.g. random vertex or edge selection, snowball sampling, the sketching-based sampling [22] etc., in order to be able to infer the properties of the original graph from the much smaller sampled graph [17, 19]. In contrast, our work constructs subgraphs of predetermined important vertices, not for the purpose of deducing properties of the original graph, but in order to infer the underlying relationships amongst the important vertices themselves.

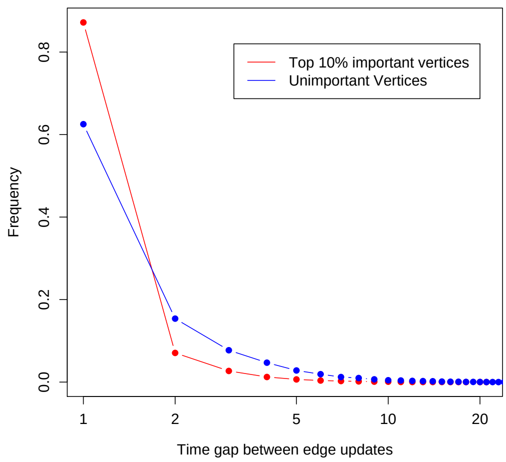


Fig. 16 The distribution of time gap between edge updates of important vertices versus unimportant vertices for the Honda forum data. The important vertices are the top 10% vertices with highest degree at the last time step.

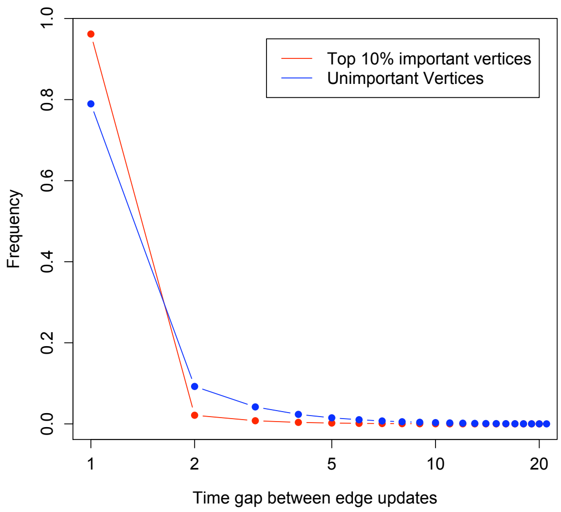


Fig. 17 The distribution of time gap between edge updates of important vertices versus unimportant vertices for the Second Life transaction data set.

In the "rich-club phenomenon",vertices with high degree tend to connect together tightly , which is true for many social and other types of real networks [37, 8]. While previous work on the rich-club phenomenon has aimed to determine whether the number of edges between high degree vertices based purely on degree is higher than what one would expect at random, our study extends to other centrality measures, and describes essential properties of the subgraphs themselves, such as connectivity, shortest paths, and preserving rank orderings of importance. A related analysis of highly interconnected sub-structures in networks is that of k-cores, subgraphs of vertices where each vertex has at least k connections within the subgraph [9]. An interesting direction for future work would be to repeat our analysis of the properties of the subgraph and original graph, using k-core membership as the importance measure for vertex selection.

8 Conclusion

In this paper, we propose a new approach to analyzing and studying large online networks, *vertex-importance graph synopsis*. Given a set of important vertices, we extract a much smaller subgraph from the original network, containing those important vertices. We attempt to place this process on a rigorous footing and show that even simple versions of the graph compression problem are hard (but that there are reasonable heuristic algorithms that can be computed relatively efficiently on real-world networks). Unlike previous methods which evaluated the fidelity of the "graph abstract," this approach utilizes the subsets of important vertices and edges and the information they could provide in large networks. These observations suggest future work in using graph synopses for information retrieval and information flow detection.

From our empirical analysis of three real online networks, we find a number of interesting properties. The important vertices are much more closely and densely connected to each other. They also have significantly shorter pairwise paths, which do not heavily depend on the rest of vertices in the networks, (i.e. their pairwise shortest paths in the subgraphs induced by themselves are close to those in the original graphs).Their relative ranks are almost all highly correlated to their ranks in the original networks. Finally, important vertices are also more persistent over time, and they acquire new edges at a higher rate.

Although our experiments show that the properties of vertices of different importance measures in different networks do vary in some ways, the observations stated above are consistent no matter the type of networks (either social or technological), and regardless of the importance measure we choose. Thus, we may use vertex-importance graph synopses as small but accurate representatives of the important vertices in the larger graph (and, sometimes, of the larger graph itself). Furthermore, the real online networks are relatively easy to compress while preserving important graph properties (they do not exhibit the worst-case behavior of our theoretical analysis).

In addition to empirical studies, we use analytical discussions to show how these properties of important vertices in online networks differ from random graph models. What is more, we also use heuristic algorithms to measure the complexities and trade-offs of requiring some properties of the real networks to be guaranteed in the compressed graphs.

9 Acknowledgments

We thank Martin Strauss for many helpful discussions. We are also grateful to Adam D'Angelo for sharing the BuddyZoo data set with us.

References

1. L. Adamic. The Small World Web. *Proceedings of ECDL*, 99:443–452, 1999.
2. L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of LinkKDD-2005*, 2005.
3. Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM Press.
4. Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of WWW '07*, pages 835–844, New York, NY, USA, 2007. ACM.
5. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
6. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.
7. F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan. On compressing social networks. In *ACM SIGKDD'09*, pages 219–228. ACM New York, NY, USA, 2009.
8. V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *NATURE PHYSICS*, 2:110, 2006.
9. S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. k-core organization of complex networks. *Physical Review Letters*, 96:040601, 2006.
10. M. Garey and D. Johnson. The rectilinear Steiner problem is NP-complete. *SIAM J. Appl. Math.*, 32(4):826–834, 1977.
11. A. C. Gilbert and K. Levchenko. Compressing network graphs. In *LinkKDD*, 2004.
12. E. N. Gilbert and H. O. Pollak. Steiner minimal trees. *SIAM J. Appl. Math.*, 16(1):1–29, 1968.
13. V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*, 2008.
14. T. Hogg and L. Adamic. Enhancing reputation mechanisms via online social networks. *Proceedings of the 5th ACM conference on Electronic commerce*, pages 236–237, 2004.
15. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
16. J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, 1627:1–17, 1999.
17. S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.
18. J. Leskovec, S. Dumais, and E. Horvitz. Web projections: learning from contextual subgraphs of the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 471–480, New York, NY, USA, 2007. ACM Press.
19. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM Press.
20. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.
21. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, 2007.
22. P. Li, K. Church, and T. Hastie. A sketch-based sampling algorithm on sparse data. Technical report, Department of Statistics, Stanford University, 2006.
23. T. Linden. The second life economy - third quarter 2009 in detail. `https://blogs.secondlife.com/community/features/blog/2009/11/02/`, 2009.
24. C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. Tech report (dcs), Dept of Computing Science, University of Glasgow, 2006.
25. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.
26. M. Newman. Coauthorship networks and patterns of scientific collaboration, 2004.
27. M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct 2002.
28. M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
29. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
30. J. Park and M. Newman. Origin of degree correlations in the Internet and other networks. *Physical Review E*, 68(2):26112, 2003.
31. R. Pastor-Satorras, A. Vazquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87:258701, 2001.
32. D. Pennock, G. Flake, S. Lawrence, E. Glover, and C. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207, 2002.
33. X. Shi, M. Bonner, L. A. Adamic, and A. C. Gilbert. The very small world of the well-connected. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 61–70, New York, NY, USA, 2008. ACM.
34. X. Shi, B. Tseng, and L. Adamic. Looking at the Blogosphere Topology through Different Lenses. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, volume 1001, page 48109, 2007.
35. X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–786, New York, NY, USA, 2009. ACM.
36. A. Y. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. In *KDD '04: Proceedings*

of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 719–724, New York, NY, USA, 2004. ACM Press.

37. S. Zhou and R. J. Mondragon. The Rich-Club Phenomenon In The Internet Topology. *IEEE Commun. Lett.*, 8:180–182, 2004.