# Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection

Fragkiskos D. Malliaros* Vasileios Megalooikonomou† Christos Faloutsos‡

**Abstract**

Given a large social graph, like a scientific collaboration network, what can we say about its robustness? Can we estimate a robustness index for a graph quickly? If the graph evolves over time, how these properties change? In this work, we are trying to answer the above questions studying the *expansion properties* of large social graphs. First, we present a measure which characterizes the robustness properties of a graph, and serves as global measure of the community structure (or lack thereof). We study how these properties change over time and we show how to spot outliers and anomalies over time. We apply our method on several diverse real networks with millions of nodes. We also show how to compute our measure efficiently by exploiting the special spectral properties of real-world networks.

**Keywords:** Social Networks; Robustness; Expansion; Measurement; Graph Mining

## 1 Introduction

Over the last few years, social networks and graphs in general, have received a considerable interest from the research community. Several kind of data arising from many diverse disciplines can be naturally represented as graphs (or networks). Characteristic examples are technological and information networks (e.g., the Web, the Internet, e-mail exchange networks), collaboration and citation networks (e.g., the DBLP co-authorship network), as well as social networks from online social networking and social media applications, like Facebook and Youtube [34]. A large amount of research work has been devoted on understanding the structure, the organization and the evolution of these networks, with many interesting results [9].

One important aspect which is related to the structure of such graphs, is the notion of *robustness*. Generally, a graph is characterized as robust, if it is capable to retain its structure and its connectivity properties after the loss of a portion of its nodes and edges. The property of robustness in real-world graphs is closely related to the notion of *community structure*. For example, consider a network with good community structure [37]. This means that the network is organized based on a modular architecture, presenting well-defined clusters with large inter-cluster and small intra-cluster edge density. We expect that the robustness of this network will be poor, since it can be easily become disconnected with the removal of the edges which connect the different clusters. How can we do this estimation quickly without removing edges and nodes and measuring the connectivity? In other words, is there a robustness and community structure index, which can be computed fast enough, even for graphs with millions of nodes and edges? Moreover, if the network evolves over time, what can we say about its robustness, and as an extension, about its community structure? Is there a common pattern in social graphs that govern the time evolution of these properties?

In this work, we tackle the problem of estimating the robustness properties of a graph quickly, providing simultaneously information about its community structure. In order to do this, we borrow concepts from the theory of expander graphs [20], and we study the *expansion properties* of several real-world time-evolving social graphs. The main contributions of this work are the following:

- *Novel robustness measure*: We propose to use the natural measure of expansion, in order to capture the robustness and the community structure of social graphs into a single number. We present how to efficiently and effectively compute this measure, exploiting the special spectral properties of real-worlds graphs.

- *Patterns of real graphs*: Applying this method to several large static social graphs, we observe that almost all these networks tend to be extremely

*Dept. of Computer Engineering and Informatics, University of Patras, Rio, Greece. Email: malliaro@ceid.upatras.gr

†Dept. of Computer Engineering and Informatics, University of Patras, Rio, Greece and DEnLab, Center for Information Science and Technology, Temple University, PA, USA. Email: vasilis@ceid.upatras.gr

‡School of Computer Science, Carnegie Mellon University, PA, USA. Email: christos@cs.cmu.edu

Table 1: Symbols and definitions.

| Symbol | Definition |
|---|---|
| $G$ | Graph representation of datasets |
| $V$, $E$ | Set of nodes and edges for graph $G$ |
| $\|V\|$, $\|E\|$ | Number of nodes and edges |
| $\mathbf{A}$ | Adjacency matrix of a graph |
| $a_{ij}$ | Entry in matrix $\mathbf{A}$ |
| $\lambda_i$ | $i$-th largest eigenvalue |
| $u_{ij}$ | $i$-th component of $j$-th eigenvector |
| $SC(i)$ | Subgraph centrality of node $i$ |
| $NSC_k(i)$ | Normalized subgraph centrality of node $i$ |
| $r_k$ | Generalized robustness index |

robust, showing good expansion properties; these findings are in accordance with previous studies about the quality of communities in large networks [27].

- *Patterns of time-evolving graphs*: We study how these properties change over time, examining the fragility evolution of real, time-evolving graphs. We observe a common pattern in the studied social graphs, as well as interesting connections with the so-called gelling-point [31].
- *Anomaly detection*: We show how to spot outliers and detect anomalies in graphs that evolve over time, examining the change of the robustness properties of the graph.

The rest of the paper is organized as follows: Section 2 gives the background. Section 3 presents the proposed method. Sections 4 and 5 present the experimental results and our observations for static and time-evolving graphs respectively. Section 6 surveys the related work and Section 7 presents the conclusions. Finally, the Appendix gives some theoretical details.

## 2 Preliminaries and Background

In this section we present the background and some preliminaries related to our approach for robustness estimation. We briefly discuss the notion of expander graphs and expansion properties and we describe their relations with the robustness and the community structure of a graph. Table 1 gives a list of used symbols with their definition.

**Expansion.** Informally, a graph is a good expander if it is simultaneously sparse and highly connected [20]. More precisely, given a graph $G = (V, E)$, the *expansion* of any subset of nodes $S \subset V$, with size at most $\frac{|V|}{2}$, is defined as the number of its neighborhood nodes (i.e., those nodes who have one endpoint inside $S$ and the other outside) over the size of the subset $S$. That is, if $N(S)$ are the neighborhood nodes of $S$, the expansion factor of the set $S$ is $\frac{|N(S)|}{|S|}$. A graph is considered to have good expansion properties if every subset of nodes has good expansion (i.e., many neighbors).

**Expansion, Robustness and Community Structure.** The study of the expansion properties of a graph can offer crucial insights about its structure; in particular they can act as a natural measure of the graph's robustness since they can inform us about the presence or not of edges which can operate as bottlenecks inside the network. Good expansion properties imply high robustness, since any subset of nodes in the graph will have a relatively large neighborhood. On the other hand, poor expansibility reflects exactly the opposite behavior. For any subset of nodes it is impossible to satisfy the constraint for a large neighborhood and hence, such kind of graphs are not robust enough, since they can be easily separated into disconnected subgraphs with the elimination of a small number of edges which connect the different subsets. If we think these subsets as cuts in a graph, good expansibility requires cuts with large size (i.e., large number of edges crossing the cut), and thus poor modularity and community structure. From the above discussion it is clear that the notion of expansion is closely related with both the robustness properties and the community structure of a graph (moreover, the expansion has been used in previous works as a quality measure for community detection and graph partitioning algorithms).

Thanks to a very well known result from the field of spectral graph theory, the expansion properties of a graph can be approximated using the spectrum of the adjacency matrix $\mathbf{A}$ of the graph [11]. More precisely, through the Alon-Milman (or Cheeger) inequality, the expansion of a graph is closely related to the *spectral gap* $\lambda_1 - \lambda_2$, i.e., the difference between the largest and the second largest eigenvalues of $\mathbf{A}$. In fact, this is a simple way for estimating the robustness of a graph: compute the spectral gap and if this is large, the graph will show good robustness, while in the opposite case the robustness will be poor. However, it is not clear how large the spectral gap should be for a graph, in order to characterize it as robust enough.

In [16] the author suggested a method for estimating the robustness of a graph, combining the spectral gap with the notion of *subgraph centrality* [14]. Generally, the subgraph centrality of a node is determined based on the number of closed walks (with odd length in order to avoid cycles in an acyclic

graph) that this node participates and it can be obtained from the spectrum of the adjacency matrix $\mathbf{A}$ as $SC(i) = \sum_{j=1}^{|V|} u_{ij}^2 \sinh(\lambda_j)$. If a graph shows good expansion properties (and thus high robustness), then $\lambda_1 \gg \lambda_2$ and for $SC(i)$, $\forall i \in V$ only the first term of the summation $(u_{i1}^2 \sinh(\lambda_1))$ will account (the contribution of the terms for $j = 2, \ldots, |V|$ will be negligible compared with that of $j = 1$). Hence, measuring the deviation from this behavior, we will be able to detect the existence (or lack thereof) of high robustness properties in a graph. This deviation can be summarized in the measure $\xi(G) = \sqrt{\frac{1}{|V|}\sum_{i=1}^{|V|}\left\{\log(u_{i1}) - \left(\log A + \frac{1}{2}\log(SC(i))\right)\right\}^2}$, where $A = \sinh^{-1/2}(\lambda_1)$ [16] (See Appendix for full justification).

However, the shortcoming of the above measure is that it is not scalable to large graphs, since it requires the computation of all the eigenvalues and their corresponding eigenvectors of the adjacency matrix $\mathbf{A}$. Moreover, it cannot be applied directly to bipartite graphs since these graphs do not contain odd length closed walks (See Appendix for more details).

## 3 Proposed Metric and Fast Estimation

While the measure presented in the previous section naturally captures the notion of robustness in a graph, it requires the computation of all eigenvalue - eigenvector pairs $(\lambda_i, \mathbf{u}_i)$, $\forall i \in V$, of the adjacency matrix $\mathbf{A}$. This becomes a computational bottleneck for large graphs with millions of nodes and edges, making the measure inefficient and practically not feasible for large scale graphs.

To overcome this problem, in what follows we present our approach for the efficient and simultaneously accurate computation of a robustness index, proposing a normalized version of the subgraph centrality together with the generalized robustness index $r_k$. The basic idea of our approach is to compute a low-rank eigendecomposition of the adjacency matrix $\mathbf{A}$, and combine it with the special spectral properties of real-world graphs.

### 3.1 Generalized Robustness Index $r_k$.

Here we present the proposed generalized robustness measure $r_k$, which can be used as a fast and scalable graph's robustness index. The motivating question behind this measure is how we can efficiently approximate the subgraph centrality of every node in the graph (Sec. 2), providing a scalable, expansion-based robustness estimation technique for large graphs, while simultaneously keeping the accuracy high.

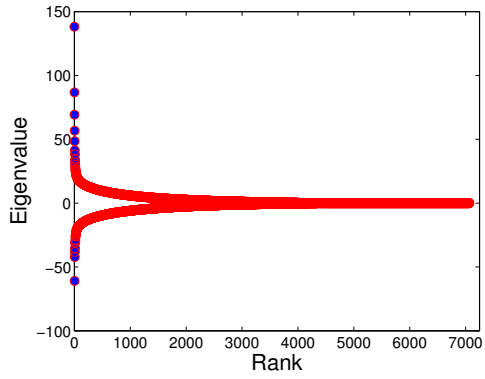



Figure 1: Skewed spectrum of a real-world network (WIKI-VOTE).

The basic idea behind our approach comes from two important observations related to the spectrum of the adjacency matrix of real-world graphs:

(i) The absolute values of the first eigenvalues follow a power-law distribution ([17]).

(ii) Except from the first few eigenvalues, the remaining eigenvalues are almost symmetric around zero, meaning that their signs tend to alternate (e.g., [41]).

Figure 1 presents the spectrum of a real-world graph (WIKI-VOTE). It plots the eigenvalues of the adjacency matrix for this graph versus their rank. We can easily observe that the first few eigenvalues are much larger than the rest and moreover the bulk of the eigenvalues are almost symmetric around zero.

Based on these points together with the fact that the $\sinh(\cdot)$ function is an odd function (i.e., $\sinh(-x) = -\sinh(x)$)[1], we can approximate the subgraph centrality (Sec. 2) using only the first top eigenvalues and their corresponding eigenvectors. In other words, the contribution of most of the eigenvalues to the subgraph centrality is negligible compared with that of the first few eigenvalues. We can now define the normalized subgraph centrality of each node in the graph as

$$NSC_k(i) = \sum_{j=1}^{k} u_{ij}^2 \sinh(\lambda_j), \ \forall i \in V \tag{3.1}$$

where $k$ is the number of the eigenvalues that will contribute to the approximation of the subgraph centrality,

[1] This simply means that the $\sinh(\cdot)$ function keeps the signs of the eigenvalues.

and generally $k \ll |V|$ for real-world graphs. In other words, $k$ can be considered as the desired low-rank approximation of the adjacency matrix $\mathbf{A}$, and as we will present in the following section, for large graphs $k$ can be extremely small to achieve almost excellent accuracy.

Based on the normalized subgraph centrality $NSC_k$ for each node $i \in N$, we can now define the proposed robustness index of a graph as

$$r_k = \left( \frac{1}{|V|} \sum_{i=1}^{|V|} \Big\{ \log(u_{i1}) - \Big( \log(\sinh^{-1/2}(\lambda_1)) + \frac{1}{2} \log(NSC_k(i)) \Big) \Big\}^2 \right)^{1/2}. \tag{3.2}$$

Smaller $r_k$ implies better robustness, since as we described in the previous section, for a robust enough graph only the first eigenpair will account for the subgraph centrality. This behavior can be visualized using the *discrepancy plot* (e.g., Fig. 4).

Definition 3.1. (Discrepancy Plot) *The log-log plot of the principal eigenvector vs. the normalized subgraph centrality will show a linear correlation for graphs with high robustness.*

Large deviation from the linear correlation in the discrepancy plot, implies absence of robustness. However, as we will see in the following section, most of the real-world social graphs we studied present this linear correlation in their discrepancy plots (as well as they exhibit a very small $r_k$ index), and therefore they tend to be extremely robust.

The $r_k$ index can be considered a generalization of $\xi(G)$ (Sec. 2) where $r_k = \xi(G)$ if $k = |V|$. However, the main advantage of the $r_k$ measure is that it is scalable and it can be computed efficiently for large graphs. Moreover, the parameter $k$ (i.e., the desirable low rank approximation) allows us to adjust the "trade-off" between the accuracy in the computation of the robustness and the required time. However, as we will present in the following section, for large graphs with millions of nodes it is enough to compute only very few of the eigenvalues and their corresponding eigenvectors to achieve almost excellent accuracy (in some cases only the first eigenvalue is adequate). The most important thing is that the $r_k$ operates perfectly as a robustness index and it can be used to summarize both the robustness and the community structure properties of a graph in a single number. Furthermore, it can be computed very easily in any programming environment that provides routines for the eigenvalue decomposition (e.g., Lanczos method [19]).

Finally we show how we can efficiently compute the $r_k$ index for bipartite graphs. Several real-world datasets can be represented as bipartite graphs. For example, consider the IMDB[2] movie-actor graph. This graph can be represented using the biadjacency matrix $\mathbf{B}$, where the rows correspond to movies while the columns to actors. A natural way to compute the robustness of this graph is to consider the actor-actor graph or the movie-movie graph (actually these graphs represent the similarity matrices between actors and movies respectively). In other words, the bipartite graph is converted into an one mode graph, projecting the nodes of one partition to the nodes of the other.

Lemma 3.1. ($NSC_k$ for bipartite graphs) *Let $\mathbf{B}_{m \times n}$ be the biadjacency matrix of a bipartite graph with $|V| = m + n$ nodes. Then, the normalized subgraph centrality $NSC_k^B$ for each node $i \in V$, can be computed as $NSC_k^B(i) = \sum_{j=1}^{k} u_{ij}^2 \sinh(\lambda_j^2)$, where $\lambda, u$ correspond to singular values and singular vectors of $\mathbf{B}$ respectively.*

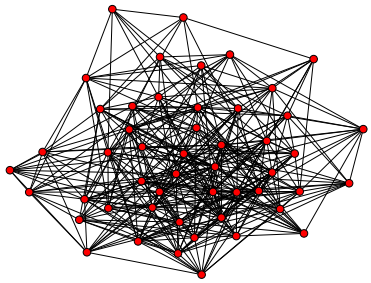
*Proof.* Applying the Singular Value Decomposition to the biadjacency matrix $\mathbf{B}$, we have $\mathbf{B}_{m \times n} = U_{m \times r} \Sigma_{r \times r} V_{r \times n}^T$. Let's consider that we represent the bipartite graph using the matrix $\mathbf{A} = \mathbf{B}^T \mathbf{B}$ (e.g., the actor-actor similarity matrix). Then, $\mathbf{A}$ is symmetric and $\mathbf{A} = V \Sigma^2 V^T$, where the matrix $V$ corresponds to the right singular vector of $\mathbf{B}$, and the eigenvalues of $\mathbf{A}$ are equal with the squares of the singular values of $\mathbf{B}$. Therefore, $NSC_k^B(i) = \sum_{j=1}^{k} u_{ij}^2 \sinh(\lambda_j^2),\ \forall i \in V$.

Thus, replacing the $NSC_k$ with $NSC_k^B$ in Eq. (3.2), we can estimate efficiently the $r_k$ index of a bipartite graph. In this paper we mainly focus on unipartite graphs, therefore we apply Eq. (3.2) as is.

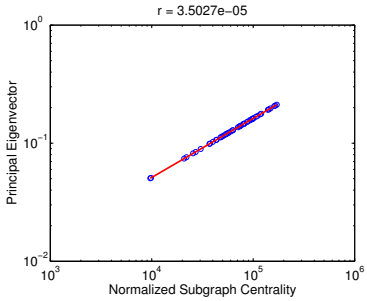
**3.2 Illustration.** In order to better understand how the $r_k$ robustness index operates, we apply it to two graphs with expected robustness properties. The first one is a random graph generated by the Erdös-Rényi (ER) model [13] with 50 nodes and probability $p = 0.3$ (Fig. 2 (a)). The second is Newman's collaboration network between 379 researchers in the area of network science (Fig. 2 (c)) [36].

Random graphs are known to have good expansion properties [20], and thus high robustness. Then, due to the large spectral gap, only the largest eigenvalue and the corresponding eigenvector will mostly contribute to the normalized subgraph centrality (Eq. (3.1)), and the principal eigenvector will follow a power-law relationship (linear correlation in logarithmic scales) with the

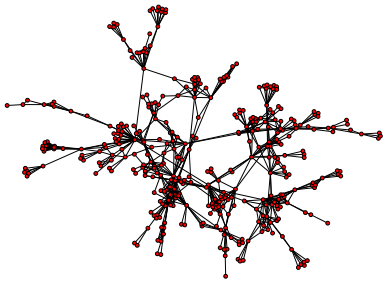
[2] www.imdb.com



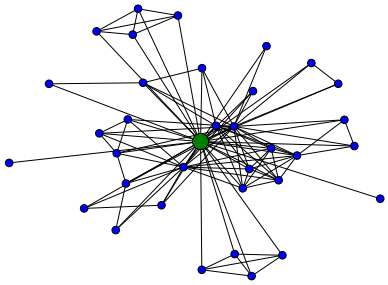
(a) ER random graph



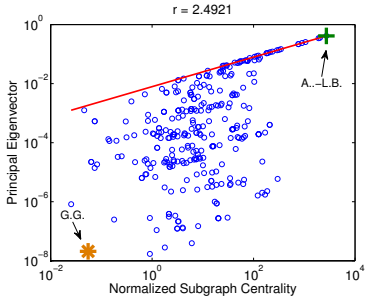

(b) Discrepancy plot



(c) Network science graph



(d) A.-L. Barabasi's egonet




(e) Discrepancy plot

Figure 2: Random vs. real graphs: Two graphs with known robustness properties and their discrepancy plots.

normalized subgraph centrality (See Appendix). Thus, from Eq. (3.2), the generalized robustness index $r_k$ will be extremely small. Figure 2 (b) depicts this result where it is easy to observe the linear correlation when plotting the principal eigenvector vs. the normalized subgraph centrality (discrepancy plot).

On the other hand, Newman's collaboration network presents very strong community structure, where the nodes form dense modules with sparse connections between different modules. Hence, this graph in not robust since it can be easily become disconnected if we simply remove the edges which connect different modules. So, we expect an opposite behavior compared with the ER graph. Figure 2 (e) depicts this result where the absence of the above linear correlation is clear in the discrepancy plot and the $r_k$ is far away from zero. Informally, the subgraph centrality measure informs us about how well clustered is a node in its neighborhood, while the principal eigenvector contains information related to the position of a node globally in the graph. So, the absence of correlation between them suggests low robustness.

Based on this, in Fig. 2 (e) the node with the largest $NSC_k$ and principal eigenvector component (green +) corresponds to A.-L. Barabasi. This is somewhat expected since A.-L. Barabasi is a well known researcher in the area of network science. Next him follow other well known researchers (e.g., H. Jeong, R. Albert) which actually belong to the egonet (Fig. 2 (d)) of A.-L. Barabasi (in Fig. 2 (d) the green node corresponds to A.-L. Barabasi). On the other hand, the node with one of the smallest $NSC_k$ and principal eigenvector (yellow $*$ in Fig. 2 (e)) corresponds to G. Gregoire, which actually has only one co-author in the dataset (and this co-author has very small neighborhood).

## 4 Robustness of Large Static Graphs

In this section we present detailed experimental results, applying the method proposed in Sec. 3 to several real-world large social graphs (Table 2). All the experiments were designed to answer the following questions:

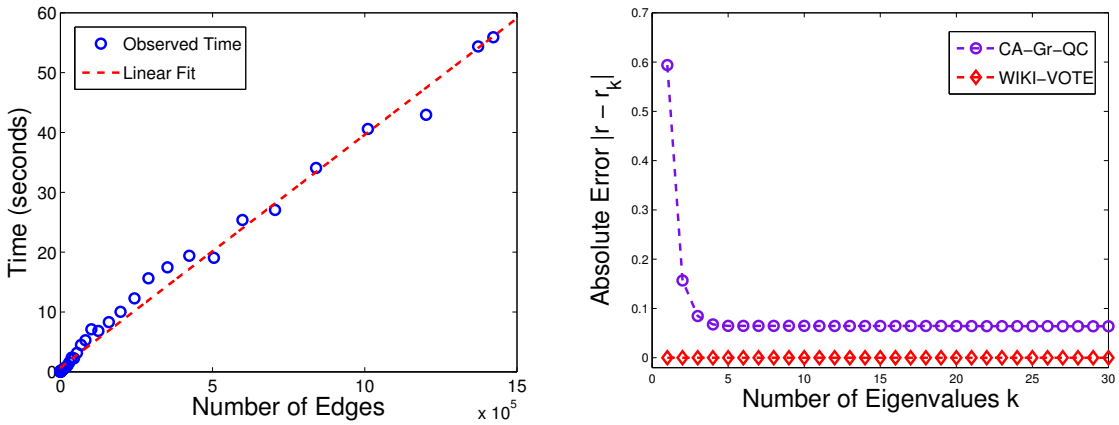
**Q1** (*Effectiveness and Scalability*) How effective and scalable (efficient) is the proposed $r_k$ index?

**Q2** (*Patterns*) What can we say about the robustness of large social graphs? Is there any common pattern that appears in most of them?

Table 2 presents the real datasets used in this work. In all cases, we consider the graphs as unweighted and undirected. Furthermore, we extract the largest connected component and use it as a good representative of the whole graph.

Table 2: Summary of real-world networks used in this study.

| Network Name | Nodes | Edges | Description |
|---|---|---|---|
| EPINIONS [38] | 75,877 | 405,739 | Who trusts whom network |
| EMAIL-EUALL [26] | 224,832 | 340,795 | Email network |
| SLASHDOT [27] | 77,360 | 546,487 | Slashdot social network (Nov. '08) |
| WIKI-VOTE [24] | 7,066 | 100,736 | Wikipedia who-votes-on-whom network |
| FACEBOOK [43] | 63,392 | 816,886 | Facebook New Orleans social network |
| YOUTUBE [32] | 1,134,890 | 2,987,624 | Social network from Youtube site |
| CA-ASTRO-PH [26] | 17,903 | 197,031 | Co-authorship network in Astro Physics |
| CA-GR-QC [26] | 4,158 | 13,428 | Co-authorship network in General Relativity |
| CA-HEP-TH [26] | 8,638 | 24,827 | Co-authorship network in High Energy Physics |
| DBLP [1] | 404,892 | 1,422,263 | Co-authorship network from DBLP |
| CIT-HEP-TH [2] | 26,084 | 334,091 | Citation network in High Energy Physics |




(a) Scalability (b) Effectiveness

Figure 3: Scalability and Effectiveness of $r_k$ index: (a) The computation time is linear with the number of edges. (b) Absolute error using $k = 1, \ldots, 30$ for two different graphs. Observe that a few eigenvalues are enough to achieve an almost excellent approximation.

**4.1 Effectiveness and Scalability of $r_k$ index.** Here we measure the performance of $r_k$ index both in terms of scalability and effectiveness. All the experiments were conducted on a DELL server with two quad core processors and 32 GB RAM, running Linux.

Figure 3 (a) presents the computation time of $r_k$ index in the DBLP dataset. In the experiment we used $k = 30$ (i.e., the 30 largest eigenpairs) and measured the running time for different scale graphs (up to $400K$ nodes and $1,4M$ edges). We can observe that the $r_k$ index scales linearly with respect to the number of edges. Moreover, we can see that for the largest graph, the computation time is less than one minute. This makes the $r_k$ index applicable to million node graphs.

Figure 3 (b) plots the rank $k$ of approximation (i.e., the number of computed eigenpairs) vs. the absolute error $|r - r_k|$, where $r$ is the value of the robustness index using the whole spectrum of the adjacency matrix, for two graphs. For the CA-GR-QC graph, we can observe that after $k = 4$ we achieve a very good approximation of the robustness index, with absolute error less than 0.06. For the WIKI-VOTE graph, for $k = 1$ and only the first eigenvalue and the corresponding eigenvector, we attain absolute error which tend to zero ($10^{-15}$). However, CA-GR-QC is a much smaller graph from WIKI-VOTE. As we will see next in this section, almost all the examined large social graphs tend to be extremely robust showing a large spectral gap, and in Eq. (3.1) the first term dominates.

**4.2 Observations and Possible Explanations.** Figure 4 presents the discrepancy plots for the graphs we examined, together with the $r_k$ index (for all the experiments we used $k = 30$). From a first look, it is clear that almost all of these graphs exhibit high robustness, showing linear correlation (in log-log scales) between the principal eigenvector and the normalized subgraph centrality. The $r_k$ index for most of them is very close to zero, implying that the spectral gap of these networks is large and they show good expansibility.

OBSERVATION 1. (HIGH ROBUSTNESS) *Large real-world social graphs exhibit good expansion properties and thus high robustness.*

This observation suggests that these networks expand very well allowing the selection of arbitrary subsets of nodes with size at most $\frac{|V|}{2}$, such that for every set there is a relatively large number of edges with one endpoint inside the set and the other outside. Therefore, a first outcome is that these social graphs lack of edges that can act as bottlenecks and thus they present high robustness. From a community structure related point of view, this observation implies that the nodes inside the networks we examined are not organized based on a clear modular architecture. It seems that these networks lack of well defined clusters which can be easily separated from the whole graph.

One interesting question is if these observations for large social graphs are expected. It is well known that the organization of social networks is based on communities (i.e., subgraphs with high intra-community and low inter-community edge density) [35]. Additionally, previous studies on the expansion properties of *small* social graphs showed that almost all of them exhibit poor expansibility and thus very low robustness [15].

On the other hand, our observations suggest an almost opposite behavior. We consider that this difference is mainly due to the scale of the networks. It seems that in large scale social graphs it is difficult to find subsets of nodes which can be easily isolated, leading to high robustness. For example, consider the co-authorship networks DBLP-1980 and DBLP-2006, in Fig. 4 (j) and (k). Both of these networks are coming from the same dataset (DBLP), but they represent different time snapshots of the graph. The DBLP-1980 graph has about $5K$ nodes and $9K$ edges, while the DBLP-2006 graph has $405K$ nodes and $1,5M$ edges. Moreover, the first graph is contained into the second. Comparing their robustness indices, it is clear that the larger network is much more robust than the smaller one. A similar argument can be used to justify the difference in robustness properties of the graphs CA-GR-QC and CA-HEP-TH (Fig. 4 (h) and (i)). As we will present in Section 5, the robustness of a graph changes over time while the graph evolves, showing interesting patterns.

Finally, our findings for static graphs are in accordance with previous works related to the quality of the community structure in large networks. In [27], Leskovec et al. observed that the best communities in large networks correspond to small subgraphs up to 100 nodes, and the quality of a community (obtained by a measure such as modularity or conductance) decreases while the size of the community increases.

## 5 Time Evolving Graphs and Anomaly Detection

In the previous section we observed that most of the studied social graphs tend to be extremely robust, presenting very low $r_k$ value. In this section we focus on time-evolving social graphs and trying to answer the following questions:

**Q3** (*Time Evolution*) How the robustness index $r_k$ of a graph changes over time?

**Q4** (*Anomaly Detection*) Can we spot anomalies over time using the $r_k$ index?

**5.1 Fragility Evolution.** As we mentioned earlier, large real-world graphs present high robustness (good expansion properties) and thus poor community structure. However, a crucial question which naturally arises for time-evolving graphs, is how these properties change over time. In order to answer these questions, we study the *fragility evolution* of a graph. In other words, for every time point in the datasets (e.g., month, year), we form the graph up to the specific time point, and then for each time snapshot we examine the $r_k$ index. We conduct experiments with the last two datasets of Table 2. DBLP covers the time period $1960 - 2006$ (cumulative graph snapshots per year) and CIT-HEP-TH expands from Feb. 1993 till April 2003 (cumulative graph snapshots per month).

Figures 5 (a) and (b) present the fragility evolution for the DBLP and the CIT-HEP-TH graph respectively. Our general observation which can be confirmed from both of these graphs is that, at the first time points, while the graphs are generally in an establishment period, $r_k$ increases gradually. This means that the graphs are not robust enough, but it seems that they exhibit good community structure. However, after a specific time point, $r_k$ starts decreasing gradually, meaning that the graphs tend to be more robust, increasing their expansion properties but loosing their community structure.

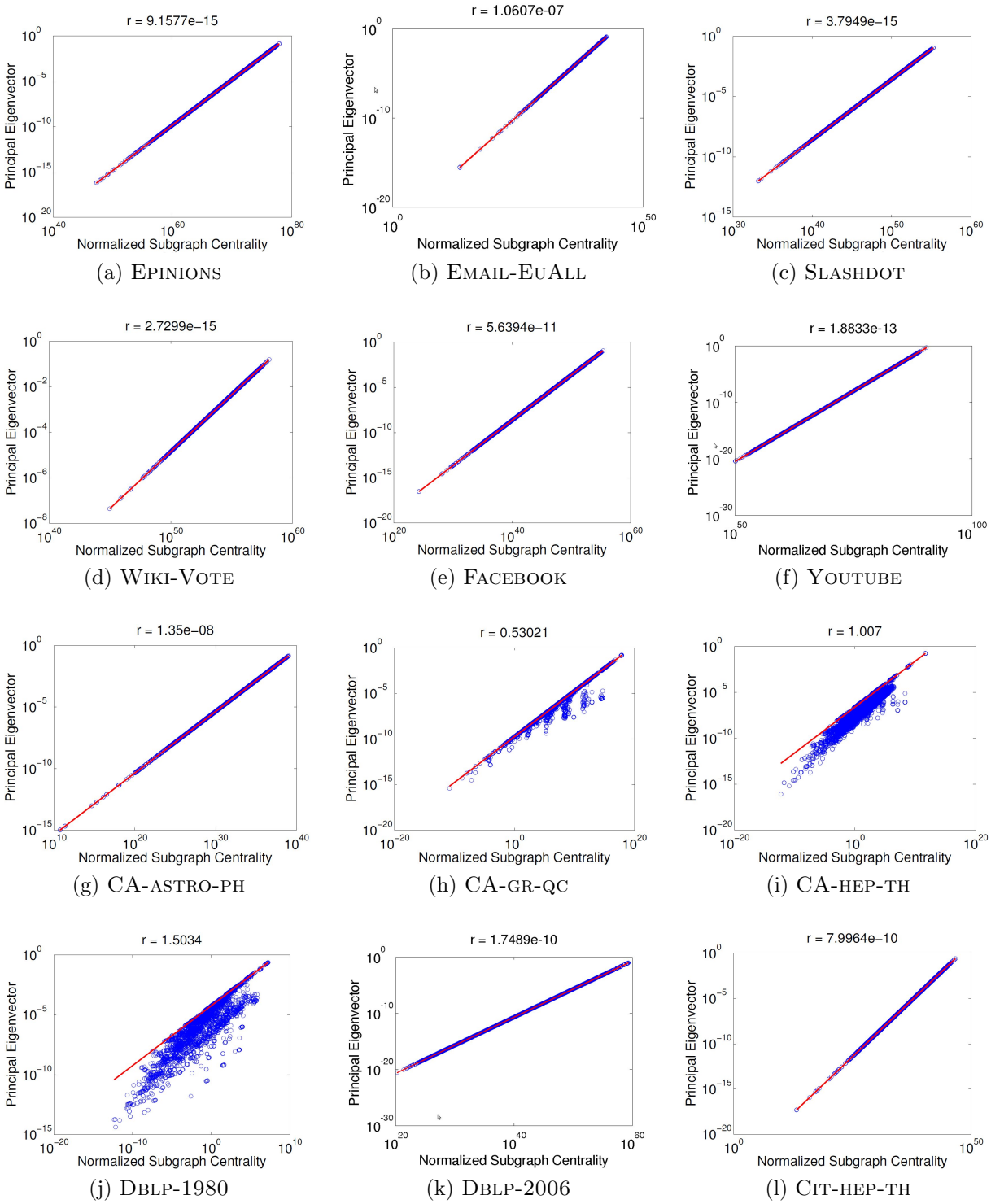
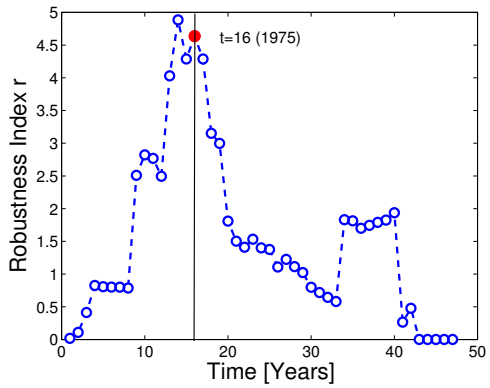
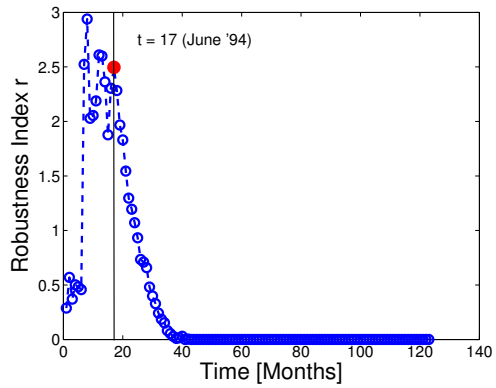


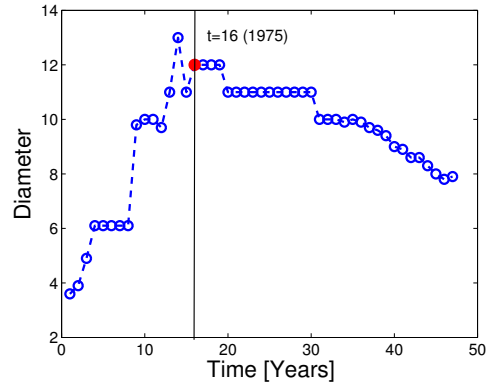

Figure 4: Discrepancy plots for several large social graphs: All plots depict the principal eigenvector vs. the normalized subgraph centrality in log-log scales, together with the $r_k$ index for each graph. Observe that almost all of them tend to be extremely robust (linearity). The red line represent the ideal behavior in case of graphs with "perfect" robustness and $r_k = 0$.



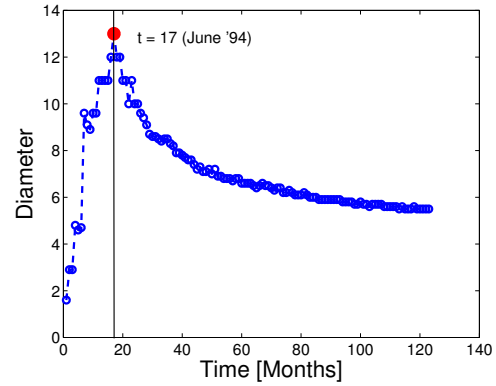

(a) DBLP: $r_k$ over time




(b) CIT-HEP-TH: $r_k$ over time




(c) DBLP: Diameter over time




(d) CIT-HEP-TH: Diameter over time

Figure 5: Fragility evolution pattern: We can observe that the spike of the $r_k$ index aligns with the diameter's spike.

Furthermore, an important thing which is related to the change of the $r_k$ index, is the time point that it occurs. We observed that this time point corresponds to the so-called *gelling point* [31]. In other words, at the time point that the graph's robustness starts improving, the effective diameter of the graph spikes (Fig. 5 (c) and (d)) and generally the graph starts obeying some of the expected rules (such as the densification law). This may be explained by the fact that there is close connection between the diameter and the robustness (expansibility) of a graph in scale-free networks [8].

OBSERVATION 2. (FRAGILITY EVOLUTION PATTERN) *Real graphs obey the fragility evolution pattern. The spike of the robustness is aligned with the gelling point.*

The fragility evolution pattern can be considered as a natural explanation for the structural differences (regarding robustness and community structure) between different scale graphs. Moreover, it seems that the $r_k$ index is an alternative way for finding the gelling point of a graph but more importantly it can be estimated more efficiently than computing the effective diameter.

**5.2 Anomaly Detection.** Here we present how the fragility evolution of a graph can be utilized for spotting outliers and detecting anomalies in graphs over time. The idea is to examine the $r_k$ index over time, trying to identify and track abrupt changes and deviations. Since for all the examined graphs presented previously the evolution of the $r_k$ index is similar, presenting a specific pattern (the fragility evolution pattern, i.e., the $r_k$ increases at the first time points and after the gelling point it starts decreasing gradually), sudden deviations from this behavior can possibly correspond to anomalies, and thus the specific time snapshots can be tagged as outliers.

Figure 6 presents the fragility evolution of the DBLP co-authorship graph (it is the same with Fig. 5 (a) but in lin-log scales).
We can observe that at two specific time points which correspond to 2002 and 2003, the $r_k$ index presents

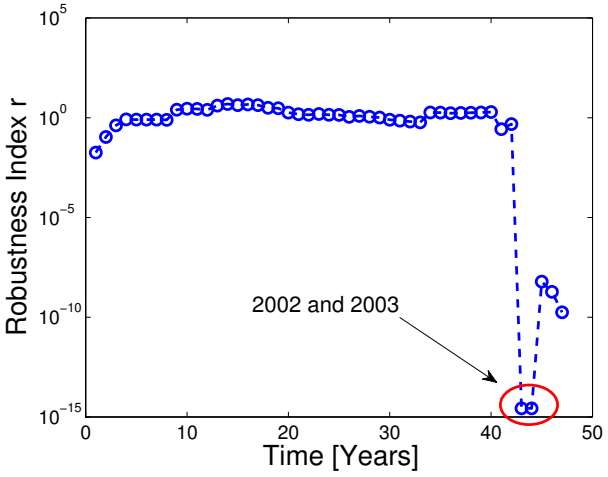



Figure 6: Fragility evolution of the DBLP graph (lin-log scales). Observe the abrupt behavior during 2002-2003. These time snapshots correspond to anomalies in the DBLP graph.

a strange behavior. More precisely, after 2001 the $r_k$ index decreases sharply and this behavior continues until 2003. After 2003 the robustness of the graph returns back to its normal behavior (it still continues to decrease but this happens gradually). These two time points present large deviation from the "normal" behavior of the graph and thus they can be classified as anomalies. In other words, it seems that for these two specific years the graph becomes extremely robust (very low $r_k$ index), but after that the robustness decreases abruptly and the graph acquires better community structure. However, are these two time graphs (2002 and 2003) really outliers, as the $r_k$ index suggests?

**Explanation.** After 2001 a large number of new publications were introduced to DBLP, which explains the downward slope of the $r_k$ index. These new publications make the co-authorship graph very robust. Until then the focus of DBLP was mostly on databases and logic programming. However, after 2002 – 2003 new research fields became important, and many old conferences and journals from these fields were added to DBLP, with focus on current publications (not in the past papers of these fields). These new fields formed new communities in the graph, decreasing the robustness, which explains the reason why $r_k$ increases after 2003. Thus, the $r_k$ index is capable to capture structural differences in the graphs and it can be used for anomaly detection in time-evolving graphs[3].

## 6 Related Work

In this section we review the related work, which can be placed into four main categories: graph structure, spectral graph analysis, applications and graph mining.

**Graph Structure.** There is a vast literature on methods for studying the structure of several kind of networks [35, 22, 32, 27, 34]. The key step for these methods is finding patterns and laws which the graphs obey. Studying static snapshots of graphs has led to the discovery of interesting properties such as the power law degree distribution [17], the small diameter [5] and the triangle power law [41]. Furthermore, examining time-evolving graphs they have been observed several patterns such as the shrinking diameter, the densification power law [25, 26] and the gelling point [31]. For a nice survey one can consult the recent work of Chakrabarti, Faloutsos, and McGlohon [9]. As far as robustness properties, Albert et al. [6] studied how scale-free networks operate under random and targeted attacks and in [16] the focus was mainly in small scale networks. In this work we focused both on fast robustness estimation without performing nodes/edges deletions, as well as on finding related patterns in both static and time-evolving large graphs.

**Spectral Graph Analysis.** Analyzing graphs using spectral techniques has a long history [11]. More recent related works include spectral algorithms for community detection [37] and spectral counting of triangles in large graphs [41, 42].

**Applications.** There are plenty of applications which involve the study of graphs. Generating realistic graphs [9] is such an application, where the generators should satisfy the observed properties. One other application which has attracted much attention is the detection of anomalies and outliers [10, 4]. Other problem domains are searching in networks [28], graph compression and summarization [29, 40, 23], graph clustering [39] and information-influence propagation in social networks [30, 7].

**Graph Mining.** A somewhat different way for studying graphs is the graph mining approach. There the main interest is focusing on mining frequent and discriminative subgraphs [44, 12, 21, 3, 18]. However, this kind of works are not directly related with our focus.

[3]Personal communication with Michael Ley and Florian Reitz from DBLP.

## 7 Conclusions

In this paper we studied the problem of estimating the robustness of social graphs, using the notion of expansion properties. The main contributions of this work are the following:

- *Fast Robustness Index*: We presented a measure which captures in a single number both the robustness as well as the community structure of a graph. We showed how to efficiently and effectively compute this measure, making it scalable for million-node graphs.
- *Patterns*: We applied the proposed $r_k$ index to several large real graphs, both static and time-evolving, and we observed the High Robustness pattern as well as the Fragility Evolution pattern.
- *Abnormality Detection:* We showed how the observed patterns related to the $r_k$ index can be used to detect anomalies in time-evolving graphs.

Future work could be the extension of the method to the MapReduce framework for studying the robustness of billion-node graphs.


## Acknowledgments

The authors would like to thank Alan Mislove and SNAP (Stanford Network Analysis Platform)[4] for providing some of the datasets used in this study, and the anonymous reviewers for valuable feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS1017415. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, the U.S. Government, or any other funding party. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.


## References


[1] `http://dblp.uni-trier.de/xml/`

[2] `http://www.cs.cornell.edu/projects/kddcup/`

[3] M. Al Hasan and M.J. Zaki. Output space sampling for graph patterns. Proc. VLDB Endow., 2 (1), pages 730-741, 2009.

[4] L. Akoglu, M. McGlohon, and C. Faloutsos. OddBall: Spotting Anomalies in Weighted Graphs. In *PAKDD*, pages 410-421, 2010.

[5] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world wide web. Nature, 401:130-131, 1999.

[6] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. Nature, 406(6794):378382, 2002.

[7] A. Anagnostopoulos, G. Brova, and E. Terzi. Peer and Authority Pressure in Information-Propagation Models. In *PKDD*, pages 76-91, 2011.

[8] B. Bollobás and O. Riordan. Robustness and Vulnerability of Scale-Free Random Graphs. Internet Mathematics, 1:1, pages 1-35, 2003.

[9] D. Chakrabarti, C. Faloutsos, and M. McGlohon. Graph mining: Laws and generators. Managing and Mining Graph Data (Chapter 3), Aggarwal, C.C., Wang, H. (eds.), Springer, 2010.

[10] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Comput. Surv., 41 (3), 2009.

[11] F.R.K. Chung. Spectral Graph Theory. CBMS, Regional Conference Series in Mathematics, No. 92, AMS, 1997.

[12] C. Chen, X. Yan, F. Zhu, and J. Han. gApprox: Mining Frequent Approximate Patterns from a Massive Network. In *ICDM*, pages 445-450, 2007.

[13] P. Erdös and A. Renyí. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5:17-61, 1960.

[14] E. Estrada and J.A. Rodríguez-Velázquez. Subgraph centrality in complex networks. Phys. Rev. E, 71, 2005.

[15] E. Estrada. Spectral scaling and good expansion properties in complex networks. Europhys. Lett., 73 (4), pages 649-655, 2006.

[16] E. Estrada. Network robustness to targeted attacks. The interplay of expansibility and degree distribution. Eur. Phys. J. B 52, pages 563-574, 2006.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *SIGCOMM*, pages 251-262, 1999.

[18] H. Fei and J. Huan. Structure Feature Selection for Graph Classification. In *CIKM*, pages 991-1000, 2008.

[19] G.H. Golub and C.F. Van Loan. Matrix computations (3rd ed.). Johns Hopkins University Press, 1996.

[20] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. Bull. Amer. Math. Soc., 43, 2006.

[21] N. Jin and W. Wang. LTS: Discriminative Subgraph Mining by Learning from Search History. In *ICDE*, pages 207-218, 2011.

[22] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, pages 611-617, 2006.

[23] K. Lefevre and E. Terzi. GraSS: Graph Structure Summarization. In *SDM*, pages 454-465, 2010.

[24] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *WWW*, pages 641-650, 2010.


[4] `http://snap.stanford.edu/`

[25] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD*, pages 177-187, 2005.
[26] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM TKDD 1 (1), 2007.
[27] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1):29-123, 2009.
[28] A.S. Maiya and T.Y. Berger-Wolf. Expansion and search in networks. In *CIKM*, pages 239-248, 2010.
[29] H. Maserrat and J. Pei. Neighbor query friendly compression of social networks. In *KDD*, pages 533-542, 2010.
[30] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. Sparsification of influence networks. In *KDD*, pages 529-537, 2011.
[31] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *KDD*, pages 524-532, 2008.
[32] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC*, pages 29-42, 2007.
[33] B. Mohar. Isoperimetric Number of Graphs. J. Comb. Theor. B 47, 274, 1989.
[34] M.E.J. Newman. The structure and function of complex networks. SIAM Review, 45:167-256, 2003.
[35] M.E.J. Newman and J. Park. Why social networks are different from other types of networks. Phys. Rev. E 68, 036122, 2003.
[36] M.E.J. Newman. Finding community structure in networks using the eigenvector of matrices. Phys. Rev. E, 74, 2006.
[37] M.E.J. Newman. Modularity and community structure in networks. PNAS, 103(23):8577-8582, 2006.
[38] M. Richardson, R. Agrawal, and P. Domingos. Trust Management for the Semantic Web. In *ISWC*, pages 351-368, 2003.
[39] V. Satuluri and S. Parthasarathy. Scalable graph clustering using stochastic flows: applications to community discovery. In *KDD*, pages 737–746, 2009.
[40] H. Toivonen, F. Zhou, A. Hartikainen and A. Hinkka. Compression of weighted graphs. In *KDD*, pages 965-973, 2011.
[41] C.E. Tsourakakis. Fast Counting of Triangles in Large Real Networks without Counting: Algorithms and Laws. In *ICDM*, pages 608-617, 2008.
[42] C.E. Tsourakakis. Counting triangles in real-world networks using projections. Knowl Inf Syst, 26:501-520, 2011.
[43] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *WOSN*, pages 37-42, 2009.
[44] X. Yan and J. Han. gSpan: Graph-Based Substructure Pattern Mining. In *ICDM*, pages 721-724, 2002.

## Appendix

Here we describe how the large spectral gap together with the subgraph centrality measure, leads to the measure $\xi(G)$ [16] as presented in Sec. 2. First of all, the subgraph centrality is defined as

$$SC(i) = \sum_{\ell=0}^{\infty} \frac{\mathbf{A}_{ii}^{\ell}}{\ell!}, \ \forall i \in V, \tag{7.3}$$

where the diagonal entry $\alpha_{ii}$ of the matrix $\mathbf{A}^{\ell}$ contains the number of walks of length $\ell$ that begin and end at the same node $i$. Focusing on unipartite graphs and keeping only the odd length closed walks[5] in order to avoid cycles in acyclic graphs, the $SC$ can be expressed as

$$SC(i) = u_{i1}^2 \sinh(\lambda_1) + \sum_{j=2}^{|V|} u_{ij}^2 \sinh(\lambda_j). \tag{7.4}$$

If the graph has good expansion properties (and thus high robustness), it means that $\lambda_1 \gg \lambda_2$, and then $u_{i1}^2 \sinh(\lambda_1) \gg \sum_{j=2}^{|V|} u_{ij}^2 \sinh(\lambda_j)$. Thus, Eq. (7.4) could be written as

$$SC(i) \approx u_{i1}^2 \sinh(\lambda_1), \ \forall i \in V. \tag{7.5}$$

This means that for graphs with high robustness, the principal eigenvector $u_{i1}$ will be related to $SC(i)$ as

$$u_{i1} \propto \sinh^{-1/2}(\lambda_1) \ SC(i)^{1/2}. \tag{7.6}$$

This relation suggests that if the graph shows high robustness, $u_{i1}$ will be proportional to $SC(i)$ and a log-log plot of $u_{i1}$ vs. $SC(i)$, $\forall i \in V$ will show a linear fit with slope 1/2 (the discrepancy plot).

[5] The bipartite graphs do not have odd length closed walks and thus the $SC$ is computed based on the even length closed walks. This happens replacing the $\sinh(\cdot)$ function with the $\cosh(\cdot)$ [14]. But then the $SC$ for the bipartite graphs cannot be efficiently approximated using similar ideas with the proposed $NSC_k$ (Sec. 3), because of the fact that the $\cosh(\cdot)$ is an even function. However, our approach for bipartite graphs (Sec. 3, Lemma 3.1) overcomes this bottleneck and can be efficiently computed for large scale graphs.