Under consideration for publication in Knowledge and Information Systems

Estimating Robustness in Large Social Graphs

Fragkiskos D. Malliaros[1], Vasileios Megalooikonomou[2], and Christos Faloutsos[3]

[1]Computer Science Laboratory, École Polytechnique, Palaiseau, France
[2]Department of Computer Engineering and Informatics, University of Patras, Rio, Greece and Center for Data Analytics and Biomedical Informatics, Temple University, PA, USA
[3]School of Computer Science, Carnegie Mellon University, PA, USA

Abstract. Given a large social graph, what can we say about its robustness? Broadly speaking, the property of robustness is crucial in real graphs, since it is related to the structural behavior of graphs to retain their connectivity properties after losing a portion of their edges/nodes. Can we estimate a robustness index for a graph quickly? Additionally, if the graph evolves over time, how this property changes?

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 also serves as global measure of the community structure (or lack thereof). We show how to compute this measure efficiently by exploiting the special spectral properties of real-world networks. We apply our method on several diverse real networks with millions of nodes, and we observe interesting properties for both static and time evolving social graphs. As an application example, we show how to spot outliers and anomalies in graphs over time. Finally, we examine how graph generating models that mimic several properties of real-world graphs, behave in terms of robustness dynamics.

Keywords: Network Robustness; Expansion Properties; Temporal Evolution; Graph Generating Models; Social Network Analysis; Graph Mining

1. Introduction

In recent years, the study of social networks and graphs in general, has received great attention from the research community. This mainly occurs due

Received Mar 18, 2014
Revised Sep 16, 2014
Accepted Dec 06, 2014

to the strong modeling capabilities that graph structures show; a large number of datasets arising from a plethora of diverse disciplines can be naturally represented as graphs. Characteristic examples are technological and information networks (e.g., the Web, the Internet, e-mail exchange networks), collaboration (e.g., co-authorship) and citation networks, as well as social networks from online social networking and social media applications, like Facebook and Youtube (Newman, 2003). A large amount of research work has been devoted to understand the structure, the organization and the evolution of these networks, with many interesting results (Chakrabarti and Faloutsos, 2012).

A cornerstone property that is related to the structure of networks, is the one of *robustness*. Broadly speaking, 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 problem of robustness assessment is one of the most well-studied in the area of network science, with many contributions from several scientific communities, including those of statistical physics and computer science (Cohen and Havlin, 2010). In the seminal paper by Albert, Jeong and Barabási (Albert et al, 2000), the robustness of real graphs was studied through a process of removing nodes and examining how some structural properties of the graph (e.g., diameter, size of largest connected component) are affected. The main observation was that the property of robustness is closely related to the degree of the removed nodes; real graphs – that present a heavy-tailed degree distribution (Faloutsos et al, 1999) – tend to be highly resilient under the removal of randomly selected nodes, while they tend to be extremely vulnerable under "targeted attacks" that focus on nodes with high degree. To conclude, the main focus of previous works was on studying graph robustness mostly based on how the removal of nodes with specific characteristics (e.g., high degree), affects structural properties of the graph.

In this paper, we argue that an important graph characteristic that plays a crucial role on the robustness, is the existence of communities. That is, 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 (Newman, 2006); this means that the network is organized based on a modular architecture, presenting well-defined clusters (i.e., communities) with large intra-cluster and small inter-cluster edge density. In other words, graphs with inherent community structure have a large number of edges between nodes of the same cluster, while relatively small number of edges across different clusters. We expect that the robustness of these types of networks will be poor, since they can be easily become disconnected with the removal of the edges which connect different clusters.

How can we do the robustness estimation quickly without performing a node/edge removal procedure and measuring how the connectivity is affected? In other words, is there a robustness index, which can be computed fast enough even for large scale graphs with millions of nodes and edges? Moreover, if the network evolves over time with the addition/deletion of nodes/edges, 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 (Hoory et al, 2006), and we study the *expansion properties* of several

real-world time-evolving social graphs. The main contributions of this work can be summarized as follows:

- *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.
- *Structural patterns of real graphs*: Applying the proposed method to several large static social graphs, we observe that almost all these networks tend to be extremely robust, showing good expansion properties. These findings are in accordance with previous studies about the quality of communities in large networks (Leskovec et al, 2009).
- *Patterns of time-evolving graphs*: We study how the robustness property of social graphs changes 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 (McGlohon et al, 2008). This pattern can be used to shed more light on the structural differences between different scale graphs.
- *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.
- *Robustness properties of graph generating models:* We study the robustness of several graph generating models and their ability to reproduce the observed patterns in static and time-evolving graphs.

The rest of the paper is organized as follows: Section 2 surveys the related work and Section 3 gives the necessary preliminary background. In Section 4 the proposed method is described. Sections 5 and 6 present the experimental results and our observations for static and time-evolving graphs respectively. In Section 7 the robustness properties of several graph generation models are examined. Some concluding remarks are presented in Section 8, and finally, the Appendix provides theoretical details.

2. Related Work

In this section we review the related work, which can be placed in the following main categories: graph structure analysis, graph robustness, spectral graph analysis and applications.

Graph Structure Analysis. There is a vast literature on methods for studying the structure of complex networks (Newman and Park, 2003; Kumar et al, 2006; Mislove, 2007; Leskovec et al, 2009; Newman, 2003). The key step for many of these approaches is the finding of patterns and laws that the graphs obey. Studying static snapshots of graphs has led to the discovery of interesting properties such as the power law degree distribution (Faloutsos et al, 1999), the small diameter (Albert et al, 1999) and the triangle power law (Tsourakakis, 2008). Furthermore, examining time-evolving graphs they have been observed several patterns such as the shrinking diameter, the densification power law (Leskovec et al, 2005; Leskovec et al, 2007) and the gelling point (McGlohon et al, 2008).

As we will present later, some of these properties are closely related to the notion of robustness that is the focus of the current work. For a nice survey one can consult the recent work by Chakrabarti and Faloutsos (Chakrabarti and Faloutsos, 2012).

Another well-known approach for exploring the structure of real graphs is the one of community detection (or clustering) (Fortunato, 2010; Malliaros and Vazirgiannis, 2013). Communities correspond to groups of nodes that tend to be similar among each other; the notion of similarity is typically expressed by the number of edges between the nodes of the same community, compared to the density of edges across different communities. As we will present later, our robustness estimation technique is closely related to the community structure property of real graphs.

Robustness Assessment in Graphs. A large number of studies has been devoted to understand and assess the robustness properties of real graphs (Cohen and Havlin, 2010). In the seminal paper by Albert et al. (Albert et al, 2000), the focus was on how scale-free networks (i.e., network that follow a heavy-tailed degree distribution) operate under random and targeted degree-based node removals. In order to assess the network robustness, one can examine how crucial structural characteristics of the graph (such as the diameter and the size of the largest connected component) behave under node removals. The main observation was that real graphs tend to be extremely robust under random failures, but vulnerable in attacks to high degree nodes. The goal of our work is slightly different; instead of performing nodes/edges removal for robustness assessment, we propose an estimation of a robustness index for a graph, based on the graph theoretic property of expansion. As we will see later, this assessment is closely related to the existence of well-defined communities within the graph. Very close to our approach is the method proposed in (Estrada, 2006); however the focus was mainly in small scale static networks, while we are interested in large real networks, as well as in the parameter of time evolution.

Spectral Graph Analysis. Analyzing graphs using spectral techniques has a long history (Chung, 1997). The main idea behind these approaches is to consider information about the spectrum of a matrix representation of the graph (mainly, the adjacency matrix or the Laplacian). More recent related works include spectral algorithms for community detection (Newman, 2006; Fortunato, 2010; Malliaros and Vazirgiannis, 2013), node centrality estimation (e.g., the PageRank algorithm (Page et al, 1999)) and spectral counting of triangles in large graphs (Tsourakakis, 2008; Tsourakakis, 2011). As we will present shortly, the proposed method can be considered as a spectral robustness estimation method, since it relies heavily on the spectrum of the adjacency matrix of a graph.

Applications. There are plenty of applications that involve the study of graphs. Generating realistic graphs (Chakrabarti and Faloutsos, 2012) is such an application, where the generators should satisfy the observed properties. As we will present later, in the context of this work we also examine how well-known graph generators behave in terms of robustness dynamics and if they can mimic the observed properties. One other application which has attracted much attention recently is the detection of anomalies and outliers (Chandola et al, 2009; Akoglu et al, 2010). Later, we will see how to utilize the observed robustness properties

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 $\mathbf{A}$
λ_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

of real graphs in order to detect temporal anomalies in social graphs. Other problem domains are searching in networks (Maiya and Berger-Wolf, 2010), graph compression and summarization (Maserrat and Pei, 2010; Toivonen et al, 2011; Lefevre and Terzi, 2010), graph clustering (Satuluri and Parthasarathy, 2009) and information-influence propagation in social networks (Mathioudakis et al, 2011; Anagnostopoulos et al, 2011).

3. Preliminaries and Background

In this section we present the necessary background and some preliminaries related to our approach for robustness estimation. Initially, we briefly discuss about the notion of expander graphs and expansion properties which form the basis of our approach and then we describe their relationship to 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 characterized as a good expander if it is simultaneously sparse and highly connected (Hoory et al, 2006). More precisely, given an undirected 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 defined as $\frac{|N(S)|}{|S|}$, and the expansion factor of the whole graph is the minimum quantity over all possible subsets S. That way, 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

Studying the expansion properties of a graph can offer crucial insights about its structure. In particular, the property of expansion can act as a natural measure of the graph's robustness since it 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 large neighborhood constraint and hence, such types 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, the existence of good expansibility requires cuts with relatively large size (i.e., large number of edges crossing the cut), and thus poor modularity and community structure (Newman, 2006).

From the discussion until now, it becomes clear that the notion of expansion is closely related to the robustness as well as to the community structure properties of a graph (note that, the expansion has been used in previous works as a quality measure for community detection and graph partitioning algorithms (Fortunato, 2010)). As we will present shortly, our approach for fast robustness estimation is built upon the above relationship; we study the robustness of a graph examining the expansion properties, providing also insights about the community structure (in other words, the property of expansion is utilized as the connecting link between robustness and community structure).

Computing the expansion factor of a graph is computationally difficult problem (Mohar, 1989). Thanks to a very well known result from the field of spectral graph theory, the expansion factor of a graph can be approximated using the spectrum of the adjacency matrix $\mathbf{A}$ of the graph (Chung, 1997). 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 the adjacency matrix $\mathbf{A}$. In fact, this constitutes 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 low. However, it is not clear *how large* the spectral gap of a graph should be in order to characterize it as robust enough. In other words, the spectral gap alone cannot provide theoretical guarantees for the expansion (and therefore for the robustness) of the real-world graphs that we are interested in (mainly graphs with heavy-tailed degree distribution). In the related literature, there have been presented some bounds for the value of spectral gap, but they mostly apply to graphs with specific properties, such as k-regular graphs (i.e., all nodes have the same degree k) (Hoory et al, 2006); however, these types of graphs and respectively the bounds, do not apply on our case.

In (Estrada, 2006) the author suggested an elegant method for estimating the robustness of a graph, combining the spectral gap with the measure of *subgraph centrality* (Estrada and Rodríguez-Velázquez, 2005). 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 the node participates, and it can be obtained from the spectrum of the adjacency matrix $\mathbf{A}$ as

$$SC(i) = \sum_{j=1}^{|V|} u_{ij}^2 \sinh(\lambda_j), \ \forall i \in V. \tag{1}$$

If the graph has good expansion properties (and thus high robustness), then due to the large spectral gap $\lambda_1 \gg \lambda_2$, we expect that in $SC(i), \ \forall i \in V$ of

Eq. (1), only the first term of the summation ($j = 1$, i.e., $u_{i1}^2 \sinh(\lambda_1)$) will account for the subgraph centrality (the contribution of the terms for $j = 2, \ldots, |V|$ will be negligible compared to that of $j = 1$ due to the effect of the $\sinh(\cdot)$ function). 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 $\xi(G) = \sqrt{\frac{1}{|V|}\sum_{i=1}^{|V|}\left\{\log(u_{i1}) - \left(\log H + \frac{1}{2}\log(SC(i))\right)\right\}^2}$, where $H = \sinh^{-1/2}(\lambda_1)$ (see Appendix for full justification) (Estrada, 2006).

However, the basic 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).

4. Proposed Robustness Measure

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. Our proposal consists of a normalized version of the subgraph centrality (Eq. (1)) along with the generalized robustness index r_k. The basic idea of our approach is to compute a low-rank eigendecomposition of the adjacency matrix $\mathbf{A}$, combining it with the special spectral properties of real-world graphs.

4.1. 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 can we efficiently approximate the subgraph centrality of Eq. (1) for every node in the graph, providing a scalable, expansion-based robustness estimation method for large graphs, while simultaneously keeping the accuracy high.

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 (Faloutsos et al, 1999).

(ii) Except from the first few eigenvalues, the remaining eigenvalues are almost symmetric around zero, meaning that their signs tend to alternate (Tsourakakis, 2008).

Figure 1 presents the spectrum of a real-world graph (Wiki-Vote). It shows 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 is almost symmetric around zero.

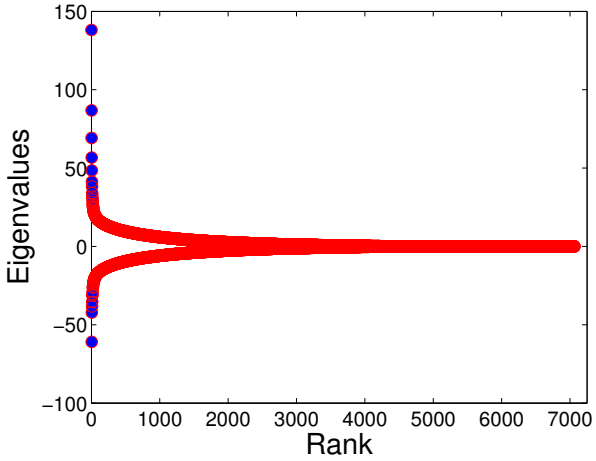


Fig. 1. Skewed spectrum of a real-world network (WIKI-VOTE). Observe that (i) there exist a few large eigenvalues, and (ii) most of the eigenvalues are almost symmetric around zero.

Based on the above two points points along with the fact that the $\sinh(\cdot)$ function, which is used to compute the subgraph centrality, is an odd function (i.e., $\sinh(-x) = -\sinh(x)$)[1], we can approximate the subgraph centrality of Eq. (1) 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 to that of the first few eigenvalues, due to the effect of the $\sinh(\cdot)$ function (the contribution of an eigenvalue is roughly canceled out by the contribution of its almost symmetric eigenvalue).

We can now define the normalized subgraph centrality of each node in the graph as

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

where k is the number of the eigenvalues that will contribute to the approximation of the subgraph centrality, 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. We stress out here that the applicability of the proposed approximation of the subgraph centrality is not similar to other low-rank matrix approximation (that retain the top-k largest eigenpairs) applications, like the ones used in other fields, such us Text Mining and Information Retrieval (e.g., Latent Semantic Indexing (Baeza-Yates and Ribeiro-Neto, 1999)). In our case, the goal is to achieve a low-rank approximation of the adjacency matrix *with respect to the subgraph centrality* function of Eq. (1). That is, the two previously described properties of real-world graphs can

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

be utilized in our approximation, only due to the specific mathematical property (odd function) of the sinh(·) function.

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

$$r_k = \left(\frac{1}{|V|} \sum_{i=1}^{|V|} \left\{ \log(u_{i1}) - \left(\log(\sinh^{-1/2}(\lambda_1)) + \frac{1}{2} \log(NSC_k(i)) \right) \right\}^2 \right)^{1/2}. \quad (3)$$

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. 2 (d) or (e)).

Definition 4.1 (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 that 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.

Let us try now to give more insights in the discrepancy plot providing a different viewpoint. Each dot in the discrepancy plot (Fig. 2 (d) or (e)) corresponds to the behavior of each individual node in the graph and shows how the node affects the robustness of the whole graph. Broadly speaking, the normalized subgraph centrality value of a node provides information about how well connected (clustered) is the node locally within its neighborhood. On the other hand, each component of the principal eigenvector captures the global connectivity of the corresponding node in the graph. Therefore, the existence (or absence) of correlations between local and global nodes' connectivity behavior, provides a good estimator for the robustness of the whole graph.

The r_k index can be considered as a generalization of $\xi(G)$ (see Section 3), 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. 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 measure 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.

4.1.1. Computational Issues

Here we discuss more technical issues related to the computation of the proposed r_k index, including time complexity and convergence in real-world graphs. As we have already mentioned, the computation of the r_k index is reduced to a symmetric eigendecomposition problem. More precisely, due to the sparsity of the adja-

Algorithm 1 Robustness Index r_k

```
function r_k = Robustness_Index (A, k)
  % Input: Adjacency matrix A, Number of eigenpairs to be used k
  % Output: The robustness index r_k
  opts.issym = 1; opts.isreal = 1;
  [u, lambda] = eigs(A, k, 'LM', opts);
  lambda = diag(lambda);
  SC = (u.^2) * sinh(lambda);
  d = log10(u(:,1)) - 0.5 * log10(SC) + 0.5 * log10(sinh(lambda(1)));
  r_k = (sum(d.^2) / length(A))^1/2 ;
end
```

cency matrix of real-world graphs, the computation of the top-k eigenpairs can be efficiently done using the Lanczos method for solving large, sparse, symmetric eigendecomposition problems. Lanczos method performs a tridiagonalization of the adjacency matrix $\mathbf{A}$, with the property that the extremal (largest or smallest) eigenvalues of the produced tridiagonal matrices approximate the extremal eigenvalues of $\mathbf{A}$. Therefore, Lanczos method is suitable in the case of computing the r_k index, where only the very few largest eigenvalues of $\mathbf{A}$ are enough. Furthermore, at each iteration of the Lanczos method, only matrix-vector multiplications are performed; hence, during the execution no intermediate matrices are produced, reducing space requirements. For a more detailed description of the Lanczos method, one can consult Ref. (Golub and Van Loan, 1996).

Another important issue concerns the convergence of Lanczos method when applied to the adjacency matrix of real graphs. As noted in (Tsourakakis, 2011), due to the skewed eigenvalue distribution of real-world networks (e.g., see Fig. 1), the convergence of Lanczos solver to the top largest eigenvalues is fast. Thus, since a relatively small number of the largest eigenpairs is adequate for computing the r_k index, this can be achieved efficiently as well.

Finally, we stress out that the proposed r_k index can be computed very easily in any programming environment that provides routines for the Lanczos eigenvalue decomposition method. For demonstration purposes, Algorithm 1 provides a MATLAB implementation of the proposed r_k index.

4.1.2. Robustness Estimation in Bipartite Graphs

Additionally, we show how we can efficiently compute the r_k index for bipartite graphs. A graph $G = (V, E)$ is called bipartite if the node set V can be partitioned into two disjoint sets V_1 and V_2, where $V = V_1 \cup V_2$, such that for each edge $(i, j) \in E \Rightarrow i \in V_1$ and $j \in V_2$. Several real-world datasets can be represented as bipartite graphs. However, the robustness index r_k cannot be applied directly to this type of graphs. The proposed optimization procedure for approximating the normalized subgraph centrality of Eq. (1), capitalizes on the combination of the skewed and almost symmetric spectrum of unipartite real-world graphs and the property that the $\sinh(\cdot)$ function is odd. As we have already mentioned, this function corresponds to the closed paths of odd length and the normalized subgraph centrality is computed according to them. However, bipartite graphs do not contain odd length closed paths; therefore, the normalized subgraph centrality should be considered based on the closed paths of even length, by replacing the $\sinh(\cdot)$ function with the $\cosh(\cdot)$ (Estrada and

Rodríguez-Velázquez, 2005). The problem that arises is that $\cosh(\cdot)$ is an even function (i.e., $\cosh(-x) = \cosh(x)$), which means that it does not retain the signs of the eigenvalues. A second problem is related to the properties of the spectrum of bipartite graphs; the eigenvalues exhibit the *pairing property*, i.e., $\lambda_{|V|-i+1} = \lambda_i,\ i = 1, 2, \ldots, |V|$ (fully symmetric around zero). Thus, it is not possible to approximate the subgraph centrality in bipartite graphs keeping only the top largest eigenvalues.

Our approach for robustness estimation performs a simple transformation of the bipartite graph in such a way that (i) the basic topological properties of the graph are retained and (ii) the subgraph centrality can be computed efficiently. As a running example, let us consider the IMDB[2] movie-actor bipartite graph (actors playing in movies – one partition corresponds to actors and the other to movies). 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 capture the similarities 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.

Proposition 4.1 (NSC_k for bipartite graphs). Let $G = (V_1, V_2, E)$ be a bipartite graph, where $|V_1| = m$ and $|V_2| = n$. Assuming that we consider the partition V_1 (similar for V_2), apply bipartite network projection procedure to construct the unipartite representation graph $\hat{G} = (V_1, E_B)$. That is, $\forall (i, w) \in E$ and $(j, w) \in E$, where $i, j \in V_1$ and $w \in V_2$ of the bipartite graph G, add the edge (i, j) in the graph $\hat{G}$. Then, the normalized subgraph centrality NSC_k^B for each node i of $\hat{G}$, can be computed as $NSC_k^B(i) = \sum_{j=1}^{k} u_{ij}^2 \sinh(\lambda_j^2)$, where now λ, u correspond to the spectrum of the adjacency matrix of $\hat{G}$.

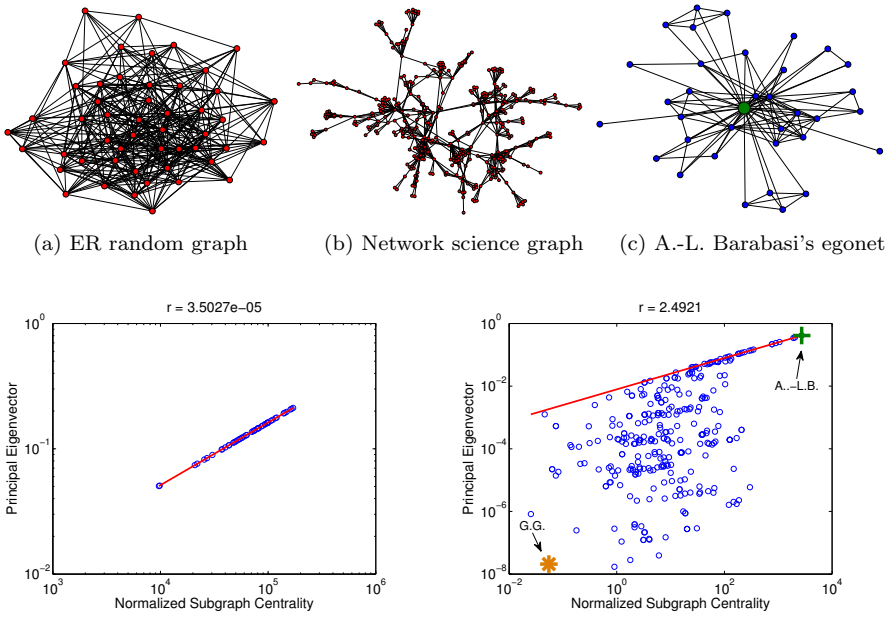
Thus, replacing the NSC_k with NSC_k^B in Eq. (3), we can estimate efficiently the r_k robustness index of a bipartite graph. In our experimental study we mainly focus on unipartite graphs, therefore we apply Eq. (3) as is.

4.2. Illustration

In order to better understand how the r_k robustness index operates, it is applied to two graphs with expected robustness properties. The first one is a random graph generated by the Erdös-Rényi (ER) model (Erdös and Rényi, 1960) 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 (b)) (Newman, 2006).

Random graphs are known to have good expansion properties (Hoory et al, 2006), 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. (2)), and the principal eigenvector will follow a power-law relationship (linear correlation in logarithmic scales) with the normalized subgraph centrality (see also Appendix). Thus, from Eq. (3), the

[2] `www.imdb.com`



(a) ER random graph (b) Network science graph (c) A.-L. Barabasi's egonet

(d) Discrepancy plot: ER random graph (e) Discrepancy plot: network science graph

Fig. 2. Random vs. real graphs and their discrepancy plots: points on the line correspond to nodes well represented by the largest "community", indicating high robustness. ER random graph ((a), (d)) is robust, while network science graph ((b), (e)) is not, consisting of several communities.

generalized robustness index r_k will be extremely small. Figure 2 (d) 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 to that of 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 index is far away from zero. As we have already discuss, low robustness is expressed by the absence of correlation between the normalized subgraph centrality (local factor) and the principal eigenvector (global factor).

Based on this point, 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 to him follow other well known researchers (e.g., H. Jeong, R. Albert) which actually belong to the egonet (Fig. 2 (c)) of A.-L. Barabasi (in Fig. 2 (c) 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).

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

Graph Dataset	Nodes	Edges
Epinions (Richardson et al, 2003)	75,877	405,739
Email-EuAll (Leskovec et al, 2007)	224,832	340,795
Slashdot (Leskovec et al, 2009)	77,360	546,487
Wiki-Vote (Leskovec et al, 2010 (b))	7,066	100,736
Facebook (Viswanath et al, 2009)	63,392	816,886
Youtube (Mislove, 2007)	1,134,890	2,987,624
CA-astro-ph (Leskovec et al, 2007)	17,903	197,031
CA-gr-qc (Leskovec et al, 2007)	4,158	13,428
CA-hep-th (Leskovec et al, 2007)	8,638	24,827
Dblp (DBLP, 2006)	404,892	1,422,263
Cit-hep-th (KDD-Cup, 2004)	26,084	334,091

5. Robustness of Large Static Graphs

In this section, we present detailed experimental results, applying the method proposed in Sec. 4 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 and Possible Explanations*) What can we say about the robustness of large social graphs? Is there any common pattern that appears in most of them?

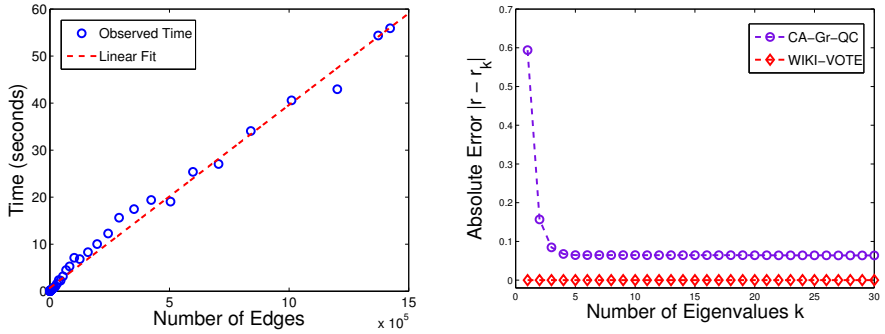
Table 2 presents the real datasets used in this work. In all cases, we consider the graphs as unweighted and undirected. Furthermore, unless specified otherwise, we extract the largest connected component and use it as a good representative of the whole graph. In Sec. 5.4 we perform experiments on the second largest connected component of several real graphs, in order to examine how the robustness behaves in smaller components.

5.1. Effectiveness and Scalability of r_k Robustness 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



(a) Scalability (b) Effectiveness

Fig. 3. Scalability and Effectiveness of r_k index: (a) The computation time is linear with respect to 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.

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 that 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. (2) the first term dominates.

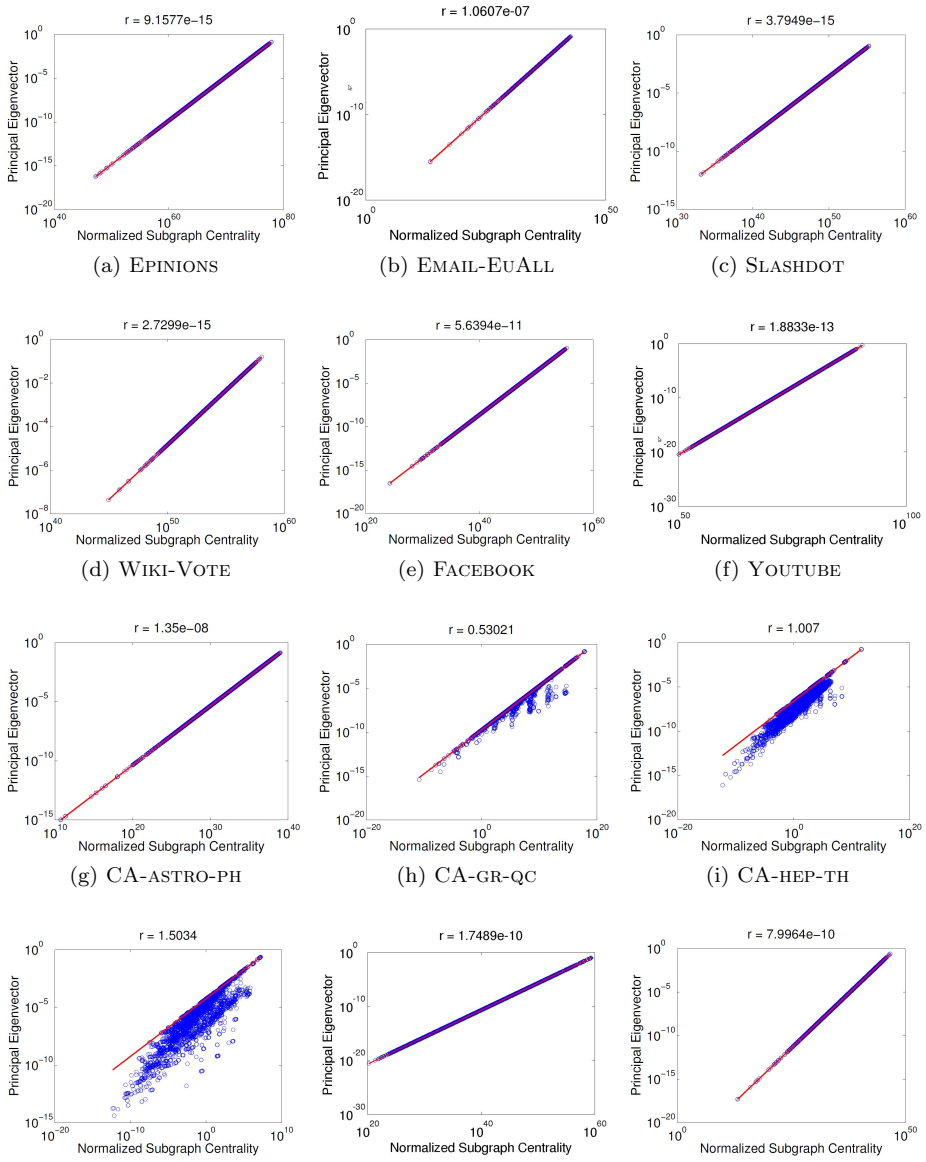
5.2. Observations and Explanations

Figure 4 presents the discrepancy plots for the graphs we examined, along 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 expansion properties.

Observation 1 (High Robustness). Large real-world social graphs exhibit good expansion properties and thus high robustness.

This observation suggests that the 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 therefore, they present high robustness. From a community structure related point of view, this observation implies that the nodes inside the networks that we have 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) (Newman and Park, 2003). Additionally, previous studies on the expansion properties of *small-scale* social graphs showed that almost all of them exhibit poor expansibility and thus very low robustness (Estrada, 2006).



(a) Epinions (b) Email-EuAll (c) Slashdot

(d) Wiki-Vote (e) Facebook (f) Youtube

(g) CA-astro-ph (h) CA-gr-qc (i) CA-hep-th

(j) Dblp-1980 (k) Dblp-2006 (l) Cit-hep-th

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

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) respectively. 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)). This is a first evidence that the robustness of a graph is not a *consistent graph property with respect to the evolution of graphs*, but as we will present in Section 6, the robustness changes over time, showing interesting patterns.

Finally, our findings for static graphs are in accordance with previous studies related to the quality of the community structure in large networks. In (Leskovec et al, 2009), the authors 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.

5.3. The Effect of Core-Periphery Structure

In a recent work, (Leskovec et al, 2009) examined the structure of several large scale social and information networks and their observation suggests that such graphs follow a *core-periphery* structure. That is, graphs are typically considered that are composed by a sparse core with no well defined structure, along with the periphery, i.e., small groups of nodes – called *whiskers* – that are barely connected to the core via a very small number of edges. Although the core does not have any clear structure, it has been shown that it is not governed by randomness, since the number of edges within core is more than the expected one (in case of random graphs). Furthermore, whiskers present diverse shapes and sizes and typically, they are organized into layers according to the number of edges used to be connected to the core. The whisker types with interesting properties are the 1-whiskers, i.e., maximal subgraphs that are connected to the core via a single edge. Leskovec et al. observed that these subgraphs are responsible for the best communities in the graph, in the sense that they achieve the best score among all subgraphs, according to some clustering quality measure (e.g., conductance).

In the experimental results presented previously, we observed that large scale social graphs tend to be highly robust. An interesting question is how the robustness is affected by the previously described core-periphery structure. Taking into consideration that in many real graphs a large portion of their nodes and edges belong to 1-whiskers (e.g., the 1-whiskers of the YOUTUBE graph include around 60% of its nodes and 24% of its edges), it is important to examine how the robustness of the graph is affected if we remove those 1-whisker nodes.

To isolate 1-whisker nodes and keep only the core of the network, we apply the method proposed in (Leskovec et al, 2009). More precisely, the core should correspond to the largest bi-connected component of the graph, i.e., to the largest subgraph in which the removal of a single edge does not affect the connectivity.

Table 3. Robustness index r_k of the initial graph compared to the one of the graph produced after removing 1-whisker nodes. The last column shows the fraction of remaining nodes and edges after removing 1-whiskers.

Graph	r_k	r_k without 1-whiskers	% of Nodes, Edges in Core
Epinions	9.1577×10^{-15}	4.7012×10^{-15}	47.59%, 90.02%
Email-EuAll	1.0607×10^{-07}	2.4706×10^{-15}	16.06%, 44.86%
Slashdot	3.7949×10^{-15}	8.6715×10^{-15}	61.04%, 97.56%
Wiki-Vote	2.7299×10^{-15}	2.4016×10^{-15}	67.73%, 97.74%
Facebook	5.6394×10^{-11}	5.4249×10^{-11}	86.19%, 98.91%
Youtube	1.8833×10^{-13}	4.3108×10^{-15}	39.83%, 76.82%
CA-astro-ph	1.3500×10^{-08}	8.8004×10^{-09}	88.97%, 98.45%
CA-gr-qc	0.5302	0.5137	63.76%, 78.12%
CA-hep-th	1.007	0.9084	68.28%, 84.69%
Dblp-1980	1.5034	1.1854	17.05%, 51.63%
Dblp-2006	1.7489×10^{-10}	1.2739×10^{-10}	69.67%, 86.68%
Cit-Hep-th	7.9964×10^{-10}	7.1770×10^{-10}	93.49%, 99.43%

Therefore, we retain only the largest bi-connected component of each graph and we compare the r_k index of this "reduced" graph to the initial one.

Table 3 shows the results of the r_k index for the graphs without 1-whisker nodes, compared to the r_k index of the original graphs. Additionally, the last column of the table depicts the fraction of nodes within the core after removing 1-whiskers. Notice that, in many cases, a relatively large portion of nodes and edges correspond to the whisker subgraphs. As we can observe, in most of the examined networks the robustness index r_k does not change significantly, meaning that the robustness of the graph is not strongly affected by these barely connected subgraphs. Moreover, the same behavior occurs in graphs with both high and low robustness (graphs in Table 3 with low and high r_k index respectively). One should expect that graphs with high robustness (like most of the graphs studied in this paper) should improve their robustness only a little after the removal of 1-whiskers. This expected behavior could be explained by the fact that whiskers correspond to sparsely connected subgraphs, that affect the expansion properties of the graph. On the other hand, it is more expected that graphs with initially low robustness and high r_k index (e.g., Ca-gr-qc, Ca-hep-th and Dblp-1980 in Table 3), should achieve a greater improvement on their robustness due to the effect of whiskers (since they are barely connected to the rest of the graph). For example, in the case of the CA-gr-qc graph in which almost 37% of its nodes and 22% of its edges belong to whiskers and were removed, the r_k index was improved by a factor of 0.03% which is relatively small. That way, the portion of the graph that finally belongs to the core is a factor that affects the robustness index. As we can see from Table 3, in graphs where only a small fraction of nodes and edges is retained after removing 1-whisker nodes (e.g., Email-EuAll and Youtube graphs), the robustness is substantially improved.

Additionally, the above behavior can possibly be explained by examining more carefully the observed core-periphery structure. As the authors in (Leskovec et al, 2009) note, the core itself has a core-periphery structure; now the periphery is composed by 2-whisker nodes that are connected via two edges to the core. Therefore, in most times, the produced graphs have a relatively small improvement on their robustness.

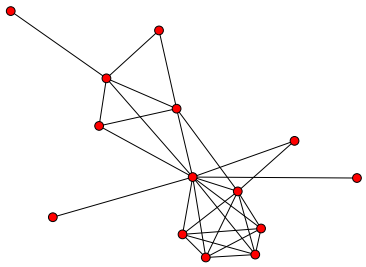
5.4. Robustness of Largest Connected Components

Most of the related works about the structure of real-world graphs focus on the Largest Connected Component (also called Giant Connected Component - GCC), i.e., the largest connected portion of the graph (e.g., see (Mislove, 2007)). To some extend, this approach can intuitively provide valuable insights about the overall structure of graphs; typically, GCCs contain the largest portion of nodes and edges and therefore they can be used as "good graph representatives" for the overall structure. However, in order to better examine and fully understand the properties of large graphs, it would be more suitable if we also study parts of the graph that do not belong to the GCC, gaining a more complete view of the underlying structure. In the related literature, only a few works have examined the structural properties of the next largest connected components (e.g., (Kumar et al, 2006), (McGlohon et al, 2008)).

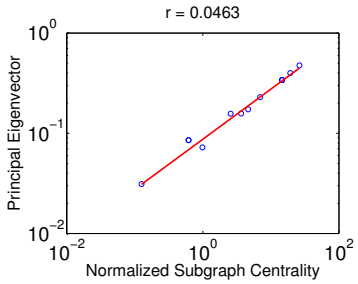
The results presented in the previous paragraphs concern the robustness of the GCC and suggest that most of the examined graphs tend to be highly robust. Here we study the robustness properties of the non Giant Connected Component of the graphs and more specifically the second Largest Connected Component (CC2); we try to examine possible deviations compared to the robustness of GCC. We performed experiments on four of the datasets presented in Table 2. Note that, in most cases the CC2 of the graphs tend to have small size (number of nodes and edges) and therefore we do not report results for extremely small CC2 subgraphs. Figure 5 depicts the discrepancy plots and a schematic representation of CC2s of the studied graphs.

First of all, the CC2s of the examined graphs follow diverse connection patterns. For example, in the YOUTUBE graph (Fig. 5 (g)), CC2 is mainly formed by a big star along with a few other nodes and edges. Regarding the robustness, almost all of the CC2 subgraphs tend to have decreased robustness (higher r_k index) compared to the one of their corresponding GCCs (Fig. 4). This can also be observed from the discrepancy plots, where there in no clear linear correlation between the normalized subgraph centrality and the principal eigenvector. Our observations suggest that the structure of CC2 is governed by different robustness patterns compared to the GCC. It seems that the structure of CC2 is more close to a modular one, in the sense that it is more easily to isolate group of nodes by deleting only a few edges (therefore lack of robustness). We consider that this observation can intuitively be explained by the temporal properties of the robustness in social graphs. Figure 6 depicts the evolution of the r_k index over time (DBLP-2006 graph), for the GCC and the CC2 respectively. As we can see, in most of the time points CC2 has almost a similar robustness behavior as the one of GCC, where in some cases (green labeled points) the component with the best robustness (GCC and CC2) is alternated. The experiments that presented in Fig. 5 concern snapshots of the CC2 subgraphs at the last time point of their evolution. As we will see shortly in Section 6, after a specific time point during the evolution of the graph (regarding the addition/removal of nodes/edges), the size of the CC2 stops increasing, while on the other hand, GCC absorbs the largest portion of nodes and edges (this phase transition point is also known as *gelling point* – marked with red color in Fig. 6). That way, CC2 and other smaller connected components constantly contain a relatively small portion of the whole graph, something that can explain the observed robustness properties of CC2.

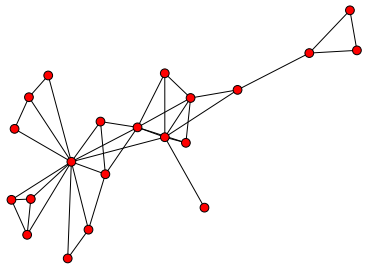
We also note that in the case of the CA-GR-QC graph, the robustness of CC2



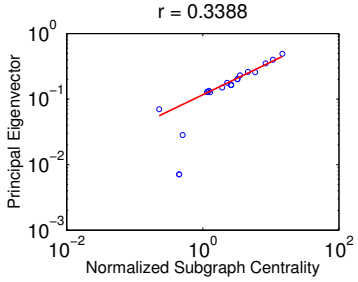
(a) CC2 of CA-GR-QC graph ...



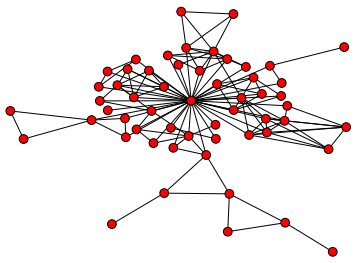
(b) ... and its discrepancy plot



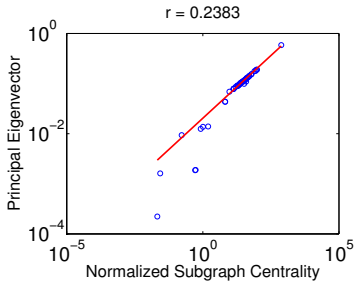
(c) CC2 of CA-HEP-TH graph ...



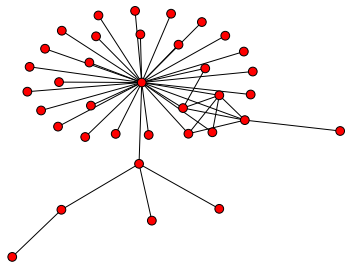
(d) ... and its discrepancy plot



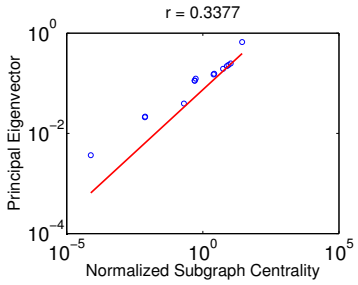
(e) CC2 of DBLP-2006 graph ...



(f) ... and its discrepancy plot



(g) CC2 of YOUTUBE graph ...



(h) ... and its discrepancy plot

Fig. 5. Schematic representation and robustness of the 2^{nd} Largest Connected Component (CC2) for several graphs.

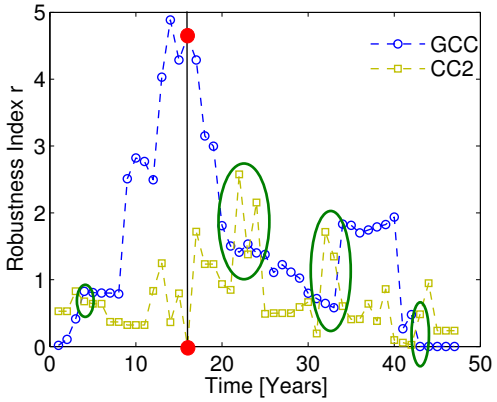


Fig. 6. Evolution of the robustness index of GCC and CC2 (DBLP-2006 graph). Observe that the robustness of CC2 does not change significantly over time.

is slightly better than the one of GCC ($r_k = 0.0463$ compared to $r_k = 0.5302$). In this case, the GCC of the graph is not robust at all, but as we can see from Fig. 5 (a), the CC2 subgraph is comprised of a relatively well-connected set of nodes.

6. Time Evolving Graphs and Anomaly Detection

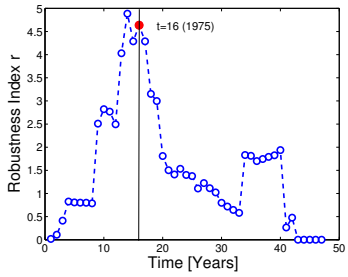
In the previous section we observed that most of the studied social graphs tend to be extremely robust, presenting very low r_k index. In this section we focus on time-evolving social graphs, in the sense that real graphs are not static but typically change over time with the addition/removal of nodes/edges (Leskovec et al, 2007). To this direction, we try 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?

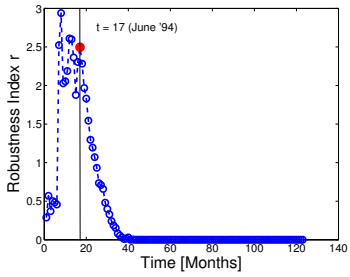
6.1. Fragility Evolution

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 February 1993 till April 2003 (cumulative graph snapshots per month).

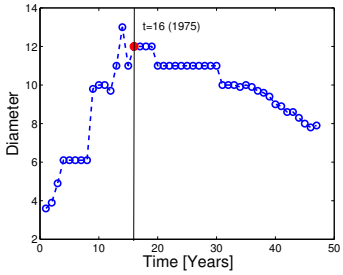
Figures 7 (a) and (b) present the fragility evolution for the DBLP and the CIT-HEP-TH graph respectively. Our general observation which can be confirmed from both of these graphs is that, at the first time points, while the graphs are



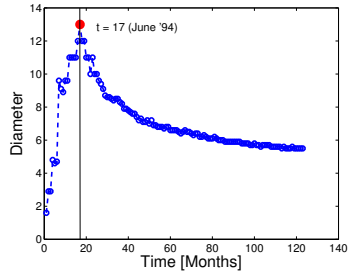
(a) DBLP: r_k over time



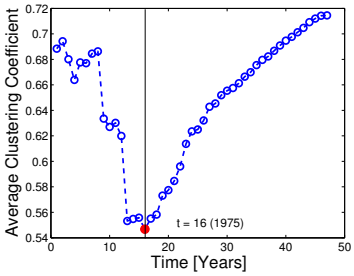
(b) CIT-HEP-TH: r_k over time



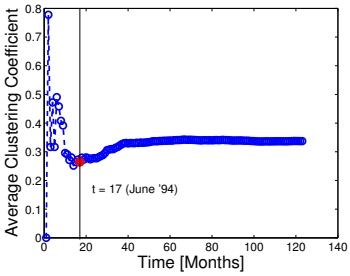
(c) DBLP: Diameter over time



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



(e) DBLP: Avg. CC over time



(f) CIT-HEP-TH: Avg. CC over time

Fig. 7. Fragility evolution pattern: We can observe that the spike of the r_k index aligns with the diameter's spike. Moreover, the average clustering coefficient (CC) over time also seems to be related to the robustness of the graph.

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.

Furthermore, an important point 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* (McGlohon et al, 2008). In other words, at the time point that the graph's robustness starts improving, the effective diameter of the graph spikes (Fig. 7 (c) and (d)) and generally the graph starts obeying some of the expected rules (such as the densification law (Leskovec et al, 2007)). This could be explained by the fact that there is close connection

between the diameter and the robustness (expansibility) of a graph in scale-free networks (Bollobás and Riordan, 2003).

Observation 2 (Fragility Evolution Pattern). Real graphs obey the fragility evolution pattern. The spike of the robustness is aligned with the gelling point.

Furthermore, we also observed a close connection of the robustness over time and the evolution of the average clustering coefficient. The local clustering coefficient (CC) (Watts and Strogatz, 1998) of each node in the graph, quantifies the extend to which its neighbors tend be connected (i.e., to create triangles). Then, the average clustering coefficient of the graph is defined as the mean value over local ones for the nodes of the graph. Watts and Strogatz used the concept of clustering coefficient to quantify the existence of small-world networks (Watts and Strogatz, 1998). As we can see from Figures 7 (e) and (f), the time point that the robustness index r_k spikes, is also aligned with the spike of the average clustering coefficient. This can be considered as an evidence that increased average clustering coefficient can lead to small world graphs with small diameter and strong robustness (a large fraction of the connected triplets within the network tend to close triangles).

The fragility evolution pattern can be considered as a natural explanation for the structural differences (regarding robustness and community structure) between different scale graphs. This means that we have a more concrete view of how the robustness (and consequently the community structure properties) change over time, and how the graph gradually improves its robustness. Finally, it seems that the r_k index is an alternative way for finding the gelling point (McGlohon et al, 2008) of a graph; more importantly it can be estimated more efficiently than computing the effective diameter.

6.2. 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 8 presents the fragility evolution of the DBLP co-authorship graph (it is the same as in Fig. 7 (a) but in linear-logarithmic scales). We can observe that at two specific time points which correspond to 2002 and 2003, the r_k index presents 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?

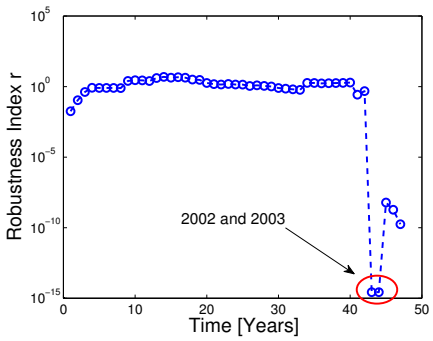


Fig. 8. Fragility evolution of the DBLP graph (lin-log scales). Observe the abrupt behavior during 2002-2003. These time snapshots correspond to anomalies in the DBLP graph.

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].

7. Robustness Analysis in Graph Generating Models

Having examined the robustness properties of several large scale social graphs, in this section we focus on investigating this property on graphs that have been produced by some well-known generating models, and we are trying to answer the following question:

Q5 (*Graph Generating Models*) How generators behave in terms of graph robustness? Do they reproduce the observed robustness patterns (on both static and time-evolving graphs)?

Answering the above questions is a quite important topic due the necessity of designing realistic graph models with properties close to the ones of real graphs. Then, the graph models can be used in a wide range of applications, such as benchmarking of graph algorithms, graph sharing (sharing of realistic but non-sensitive graphs produced by a model) and graph evolution (e.g., how the Facebook social graph should look like in one year from now?). Furthermore, a particularly significant point is that graph generators can provide useful insights about the underlying generative process of real-world graphs, towards a better and more clear understanding of their structure and formation dynamics.

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

In this work we focus our attention on three well-known and widely used graph generation modes, namely the Preferential Attachment model (Barabási and Albert, 1999), the Kronecker Graph model (Leskovec et al, 2010 (a)) and the Forest Fire model (Leskovec et al, 2005), (Leskovec et al, 2007). Although all these models do not capture exactly the full set of properties of real graphs, they all produce heavy-tailed degree distributions and are good representatives of most other models. We also note that we do not perform experiments with the Erdös-Rényi random graph model (Erdös and Rényi, 1960), as its properties are far away from those of real graphs (Chakrabarti and Faloutsos, 2012).

Furthermore, we are interested in both a *qualitative* and *quantitative* study of the robustness behavior of graph models. That is, we want to examine to what extend the models can reproduce qualitatively the observed properties in general, i.e., the high robustness and the fragility evolution patterns as trends of real graphs, as well as their ability to capture (fit) the observed r_k index of real graphs. Regarding the last point, a direct comparison between the r_k indices would not be appropriate due to the nature of the r_k index; we argue that by comparing the order of magnitude of r_k we can draw more useful conclusions.

7.1. Preferential Attachment Model

The Preferential Attachment (PA) model was introduced by Barabási and Albert (Barabási and Albert, 1999) to capture the heavy-tailed degree distribution of real-world graphs. The model operates similar to a *rich-gets-richer* mechanism, in the sense that high degree nodes are more probable to increase their degree during the evolution process. More precisely, the PA model can be described by two parameters: (i) the number of nodes n in the produced graph and (ii) the number of edges m created by each new node introduced in the graph. At every iteration step, a new node is added to the graph. The node forms m edges by connecting to m already existing nodes preferentially, i.e., each of the m endpoint nodes is selected with probability proportional to its degree. It has been shown that the PA network model, among other properties, presents a heavy-tailed degree distribution. However, as it have been discussed in the related literature, the PA model is not able to reproduce important temporal properties of real graphs, namely the densification power law and the shrinking diameter.

Fitting Process. Let us now discuss how can we set the two parameters of the PA model (n, m) in order to generate graphs with properties (i.e., number of nodes and edges) similar to those presented in Table 2. The number of nodes n is set equal to the number of nodes in the original graph $|V|$. Since each of the $n = |V|$ nodes in the network creates m edges, the total number of edges would be $|E| = m \cdot |V|$. Therefore, parameter m is set to be equal to the density factor $m = \lceil |E|/|V| \rceil$.

In general, we know that PA graphs are robust in random failures, but tend to be vulnerable under targeted attacks to high degree nodes (Albert et al, 2000). Here we examine how the r_k robustness index behaves and if the PA model is capable to reproduce the high robustness (see Section 5) and the fragility evolution patterns (see Section 6). Table 4 shows a quantitative comparison between the r_k robustness index of the original graph and the one of the artificial graphs produced by the PA model, following the previously described fitting

Table 4. Robustness of graphs generated by the Preferential Attachment model. Parameter m represents the number of connections that each new node creates.

		Preferential Attachment	
Graph	r_k	m	r_k
EPINIONS	9.1577×10^{-15}	5	5.7053×10^{-05}
EMAIL-EUALL	1.0607×10^{-07}	2	0.0915
SLASHDOT	3.7949×10^{-15}	7	6.7911×10^{-07}
WIKI-VOTE	2.7299×10^{-15}	14	1.7554×10^{-15}
FACEBOOK	5.6393×10^{-11}	13	2.4337×10^{-15}
YOUTUBE	1.8833×10^{-13}	3	0.0642
CA-ASTRO-PH	1.3500×10^{-08}	11	3.3600×10^{-11}
CA-GR-QC	0.5302	3	0.0051
CA-HEP-TH	1.0070	3	0.0070
DBLP-1980	1.5034	2	0.0604
DBLP-2006	1.7489×10^{-10}	4	0.0214
CIT-HEP-TH	7.9964×10^{-10}	13	1.4641×10^{-15}

process. For every graph we also show the value of parameter m of the PA model.

Comparing the r_k index with that of Fig. 4 (also second column of Table 4), we can observe that in most cases the PA model fails to reproduce the robustness properties of the graphs. Although in some cases it seems that the r_k index is close enough to the original one (e.g., in the WIKI-VOTE graph), we consider that this mainly happens due to the settings of parameter m (i.e., $m = \frac{|E|}{|V|}$). That is, the r_k value of each PA graph, depends heavily on the number of edges that each incoming node creates, i.e., parameter m of the model.

To provide a more thorough examination of this behavior, we study the robustness index of PA graphs at different scales and for several values of parameter m. Figure 9 depicts the r_k value for PA graphs of various sizes ($5K - 350K$ nodes), and for a wide range of values for parameter m ($m = [2, 4, 6, 8]$). We can observe that the PA model fails to qualitatively reproduce the fragility evolution pattern (r_k initially increases and after a time point it starts decreasing gradually leading to robust enough graphs). More precisely, the first observation is that for small values of m (i.e., $m = [2, 4]$), the robustness seems to remain almost constant, *independently* of the graph size. This can be explained by the structural properties of the graphs generated by the PA model. In fact, it has been shown that the PA model produces graphs that exhibit constant conductance and spectral gap (Mihail et al, 2003). Thus, the robustness of the graph, as captured based on the notion of spectral gap, will also be constant as the graph evolves – which is not the case of real social graphs. We also noticed that, as parameter m increases, the r_k index presents a small upward trend with respect to the size of the graph.

Furthermore, for larger values of m, the robustness is improved (r_k index decreases); this can be considered as an evidence that if each new node creates more edges in the graph, the overall connectivity of the graph potentially will increase. However, we can observe that even for the largest examined value of $m = 8$ and the largest graph of $|V| = 350K$ nodes, the r_k index cannot reach extremely low values similar to the ones of real-world graphs. As we can see

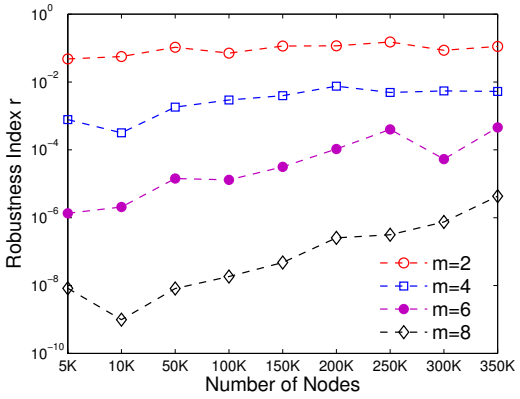


Fig. 9. Robustness index of the Preferential Attachment (PA) graph model. Observe that even for large values of parameter m and graph size $|V|$, the PA graphs cannot qualitatively reproduce the observed robustness patterns.

from Table 4, this can only be achieved for very large values of m (e.g., as the WIKIVOTE graph).

To conclude, we observed that the PA model cannot mimic the robustness evolution of real graphs, mostly due to the constant conductance and spectral gap properties. Definitely, the PA model does not reproduce several other recently observed properties of real-world graphs, such as the shrinking diameter (Leskovec et al, 2007).

7.2. Kronecker Model

The second model that we have examined in our study, is the Kronecker graph model and the ability of Kronecker graphs to reproduce the observed robustness patterns. The Kronecker model was introduced by Leskovec et al. (Leskovec et al, 2010 (a)) as a simple generation model for real-world graphs, based on the Kronecker product of matrices. More precisely, assuming an initiator adjacency matrix $\mathbf{A}_1$ of size $\ell \times \ell$, the Kronecker graph after k iterations is defined as the graph with the following adjacency matrix:

$$\mathbf{A}_k = \underbrace{\mathbf{A}_1 \otimes \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_1}_{k \text{ iterations}} = \mathbf{A}_{k-1} \otimes \mathbf{A}_1. \tag{4}$$

In practice, a stochastic version of the Kronecker model is used, in the sense that the initiator matrix $\mathbf{A}_1$ is not the binary adjacency matrix itself but the probability matrix for the existence of an edge. For example, in the typical case of a 2×2 initiator matrix $\mathbf{A}_1 = [a\ b; c\ d]$, each value represents the probability of existence of the corresponding edge. Starting by such an initiator matrix and applying the Kronecker product for a desired number of iterations k, the resulting adjacency matrix of the graph corresponds to a realization of the matrix $\mathbf{A}_k$, i.e., each edge (i, j) is introduced to the graph with probability $A_k(i, j)$. For the rest of our presentation, by the term Kronecker Graph we will refer to the stochastic version of the model.

Table 5. Robustness properties of Kronecker graphs for 2×2 and 3×3 initiator matrices respectively, compared to the r_k index of the original graphs.

		Kronecker Graphs	
Graph	r_k	r_k (2×2 Init. Matrix)	r_k (3×3 Init. Matrix)
Epinions	9.1577×10^{-15}	2.6215×10^{-07}	1.6437×10^{-12}
Email-EuAll	1.0607×10^{-07}	0.0517	0.2071
Slashdot	3.7949×10^{-15}	6.6933×10^{-07}	1.6842×10^{-09}
Wiki-Vote	2.7299×10^{-15}	2.1923×10^{-15}	2.3512×10^{-09}
Facebook	5.6393×10^{-11}	1.8354×10^{-15}	3.2705×10^{-08}
Youtube	1.8833×10^{-13}	0.0474	0.0713
CA-astro-ph	1.3500×10^{-08}	4.5312×10^{-07}	1.4090×10^{-09}
CA-gr-qc	0.5302	0.3681	0.3673
CA-hep-th	1.007	0.0932	0.4170
Dblp-1980	1.5034	0.2683	0.4599
Dblp-2006	1.7489×10^{-10}	0.0641	0.0111
Cit-Hep-th	7.9964×10^{-10}	2.4143×10^{-15}	9.7254×10^{-16}

Kronecker graphs (and some very recent extensions (Seshadhri et al, 2013)) have been proved to produce graphs with properties similar to those of real graphs. More specifically, the Kronecker graphs capture two well-known properties of dynamic graphs, namely the densification power-law and the property of shrinking diameter (Leskovec et al, 2007). Our goal is to investigate the ability of Kronecker graphs to reproduce the observed robustness properties. Next, we briefly describe how the fit a Kronecker graph to a real network, i.e., how to choose the parameters of the initiator matrix $\mathbf{A}_1$.

Fitting Process. As we mentioned earlier, the Kronecker graph model can be described by the $\ell \times \ell$ initiator matrix $\mathbf{A}_1$. In order to decide the values of the initiator matrix that could finally lead to a graph structurally similar to the target one (after applying the Kronecker product for a number of iterations), we apply the KronFit algorithm presented in (Leskovec et al, 2010 (a)). KronFit is based on maximum likelihood and sampling techniques to estimate the parameters of the initiator matrix. Furthermore, the size ℓ of the initiator matrix $\mathbf{A}_1$ is also a parameter of KronFit. We performed experiments for 2×2 and 3×3 initiator matrices. After computing $\mathbf{A}_1$ for each of the graphs presented in Table 2, we apply the Kronecker product of Eq. (4) for k iterations, until we reach the number of nodes of the target graph.

Let us now provide a quantitative discussion of the robustness properties of the Kronecker graphs and how close they are to the ones of real graphs. Table 5 shows the robustness index r_k of the graphs generated by the Kronecker model following the above fitting procedure, for two different sizes of the initiator matrices, namely 2×2 and 3×3. For comparison reasons, the second column presents the r_k index of the original graphs (as described in the experiments of Section 5). As we can observe, for some of the examined graphs, the robustness of the artificial Kronecker graphs is relatively close to the one of real graphs. Note that, we do not perform absolute comparison between the original and reproduced r_k values. In such a case, the absolute or relative error of r_k values for most of the studied graphs, should indicate that the reproduced values are far away from the original ones. A direct comparison of the indices would not be appropriate due

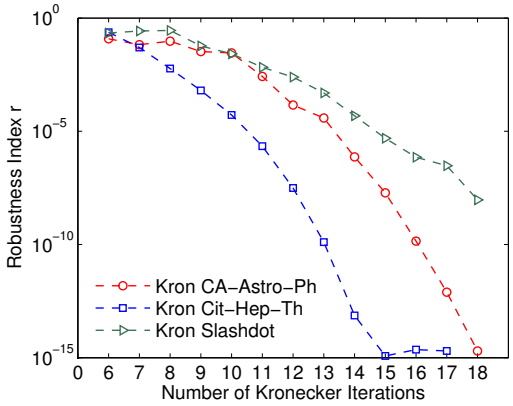


Fig. 10. Robustness index of Kronecker graphs (2×2 initiator matrix) for several iterations of the Kronecker product.

to the fact that, during the iterations of the Kronecker product, the size (number of nodes and edges) of the produced graph does not match exactly to the size of the original one. Nevertheless, we are mostly interested to qualitatively examine if the Kronecker model can capture the existence or not of robustness in graphs, and as the results suggest, there are cases of graphs (of both high and low robustness) that the model performs relatively well. However, in a few graphs (EMAIL-EUALL, YOUTUBE and DBLP-2006), the Kronecker models fails to reproduce the observed robustness properties (even at a qualitative level).

Regarding the fitting capabilities of 2×2 and 3×3 initiator matrices, both of them perform qualitatively similar. However, in many cases, the r_k values produced by these different size matrices, deviate. This mainly happens due to the way that the Kronecker product of matrices (or graphs) works. As we mentioned above, it is quite difficult to match the exact size of the graph; at every iteration of the Kronecker product, the size of the corresponding graph is raised to a power that depends on the size of the initiator matrix (square and cubic respectively).

Furthermore, we have examined the ability of Kronecker model to qualitatively reproduce the fragility evolution pattern, i.e., how the robustness of the graph changes while the graph evolves over time. More precisely, for three graphs of Table 2, we have generated the Kronecker graphs beginning from the corresponding 2×2 initiator matrices produced by the KRONFIT method. For every iteration of the Kronecker product, we have computed the r_k index and the results are presented in Fig. 10.

As we can observe from Fig. 10, the robustness of the generated Kronecker graphs improves gradually with respect to the number of Kronecker iterations (r_k index decreases). This observation suggests that the Kronecker graphs can qualitatively capture the evolution of the robustness properties of real graphs, reproducing the fragility evolution pattern. In other words, the Kronecker product can potentially produce graphs with ever increasing robustness (until some point). Of course, different settings for the initiator matrix will cause a slight deviation to the form of the fragility evolution pattern, as the different curves in Fig. 10 suggest.

We should note here that in some cases and depending on the graph dataset, we observed that the Kronecker model cannot capture exactly the small upward trend on the fragility evolution pattern (i.e., the time points where the robustness seems to decrease since the graph is still in an establishment period). This mainly happens due to the fact that the size of the produced graphs (at every Kronecker iteration) is increasing quickly due to the effect of the Kronecker product. Therefore, only for a very small number of the very initial snapshots, the robustness of the graph tends to be low (e.g., the first three iterations of the Kron Slashdot graph in Fig. 10).

To conclude, we found out that Kronecker graphs are able to capture the temporal evolution of the robustness index. However, the KronFit process, in many cases, cannot reproduce the actual robustness indices of static graphs.

7.3. Forest Fire Model

The last graph generating model that we have examined is the Forest Fire (FF) model (Leskovec et al, 2005), (Leskovec et al, 2007). The basic idea here is to design a mechanism based on which new-coming nodes $v_i, i = 1, \ldots, k$ are attached to existing nodes of the graph G, in such a way that the resulting graph will obey important properties, such as heavy-tailed degree distribution, the densification power law and shrinking diameter. At each time step t of the model, a node v that is entering in the network, chooses an ambassador node w randomly and creates a link to w. Then, based on two parameters, namely the *forward burning probability* p_f and the *backward burning probability* p_b, node v selects a subset of w's neighbors to create out-going and in-coming edges respectively. This last step is recursively applied to all v's new neighborhood nodes.

Fitting Process. Here we describe how to fit the parameters of the FF model in order to generate graphs that are close the those of Table 2. Since in this work we are interested in undirected networks, we consider only the forward burning probability p_f, and following the discussion in (Leskovec et al, 2007) we set $p_b = 0.32$ for the backward burning probability. Thus, given as input the number of nodes $|V|$ of the original graph, we search the parameter space of p_f using step of $\delta = 0.001$, in order to generate a graph with number of edges $|E|$ close to the real one. Table 6 gives the values of p_f for each of the examined dataset.

Next, we examine how the graphs generated by the FF model behave in terms robustness dynamics. Table 6 shows the robustness index r_k for the FF synthetic graphs, after selecting appropriate values for the forward burning probability p_f in order to fit the size with that of the original graph (for comparison reasons, we also provide the r_k index of the original graph). We can observe that for most of the examined graphs, the robustness of the FF artificial graphs is relatively close to the real one. In other words, the FF model is able to *qualitatively* reproduce the observed robustness r_k, both in graphs with extremely small r_k (e.g., Wiki-Vote dataset) as well as in graphs with large r_k (e.g., CA-hep-th graph). In many cases (e.g., Slashdot graph), the r_k index of the generated graphs is almost the same to the one of the real networks, while in some other cases, although the generator still produces graphs with small r_k index, the absolute values are not very close (e.g., CA-astro-ph graph). Again, we stress out here

Table 6. Robustness of graphs generated by the Forest Fire model. The backward burning probability is fixed to $p_b = 0.32$.

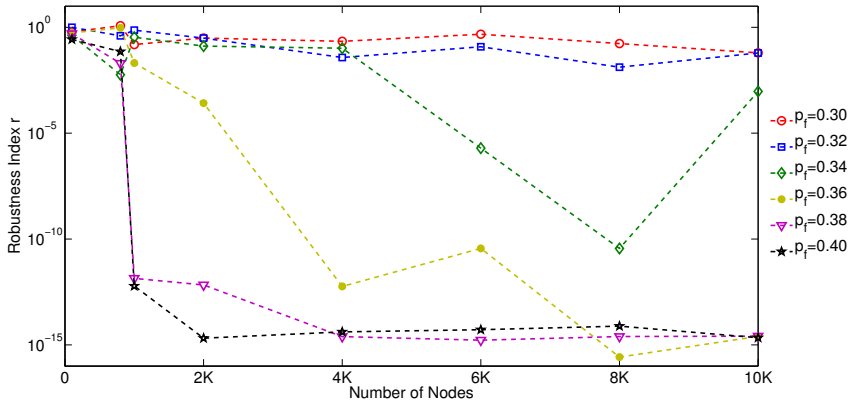
		Forest Fire	
Graph	r_k	p_f	r_k
EPINIONS	9.1577×10^{-15}	0.349	2.4826×10^{-15}
EMAIL-EUALL	1.0607×10^{-07}	0.119	4.4598
SLASHDOT	3.7949×10^{-15}	0.354	3.6953×10^{-15}
WIKI-VOTE	2.7299×10^{-15}	0.386	4.2643×10^{-15}
FACEBOOK	5.6393×10^{-11}	0.366	5.6253×10^{-15}
YOUTUBE	1.8833×10^{-13}	0.295	6.5224×10^{-05}
CA-ASTRO-PH	1.3500×10^{-08}	0.372	5.2723×10^{-15}
CA-GR-QC	0.5302	0.337	0.1114
CA-HEP-TH	1.0070	0.320	0.0974
DBLP-1980	1.5034	0.241	1.6278
DBLP-2006	1.7489×10^{-10}	0.325	5.2084×10^{-14}
CIT-HEP-TH	7.9964×10^{-10}	0.372	8.1218×10^{-15}

that an absolute comparison of these values would not be fair due to the nature of the r_k index. Nevertheless, we can infer that the FF model has the ability to generate graphs with robustness properties close to the real ones.

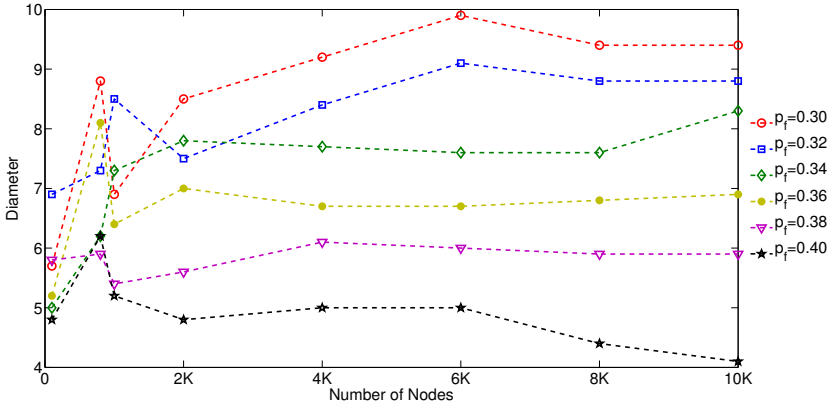
An interesting point here is the behavior of the EMAIL-EUALL graph. In this case, the FF model fails to reproduce the observed r_k index, where the produced graphs tend to be extremely non robust ($r_k = 4.4598$). One possible explanation of this observation can be derived by the settings of the forward probability value, i.e., $p_f = 0.119$, which tends to be away from most of the p_f values for the other datasets (p_f values are roughly in the range $[0.25, 0.39]$, as shown in Table 6). That is, in order to match the number of edges of the targeted graph, the value of p_f is set relatively low. According to (Leskovec et al, 2007), this forward probability value lead to graphs with small densification factor, something that explains the observed property. Note that, the Kronecker model presented earlier, also fails to reproduce the observed property for this graph dataset.

We have also examined the ability of FF model to reproduce the way that the property of robustness changes as the graph evolves over time (fragility evolution pattern). FF has been shown to produce graphs that densify over time and their diameter shrinks – which is the case of real-networks. For values of the forward burning probability p_f in the range $[0.3, 0.4]$ and different number of nodes $(100 - 10K)$, we study the behavior of the r_k index. That way, for each different value of p_f, we consider that we have a graph that evolves over time based on the properties of the FF model and we are interested to examine the temporal evolution of the r_k index.

Figure 11 (a) depicts the evolution of r_k under the FF model. As we can observe, different values of p_f typically lead to different behavior of the r_k index, as the graph grows (note that, for small graphs, e.g., less than $1K$ nodes, there is no clear difference in the behavior of the robustness). For small values of p_f, i.e., $p_f = 0.3$ and $p_f = 0.32$, the r_k index is almost constant as the size of the graph increases. This behavior clearly deviates from the fragility evolution pattern observed in real graphs. To gain further insights about this behavior, we have also examined the evolution of the effective diameter under the FF model, for these values of p_f. As we can see from Fig. 11 (b), in the case of $p_f = 0.3$



(a) Robustness index r_k



(b) Diameter

Fig. 11. (a) Robustness index of the Forest Fire graph generator and (b) the effective diameter of the corresponding graphs, for various graph sizes and values of the forward burning probability p_f.

and $p_f = 0.32$ the diameter increases over time, something that deviates from the shrinking diameter pattern of real networks (Leskovec et al, 2007).

The rest values of p_f present interesting behavior. When $p_f = 0.34$ and $p_f = 0.36$, the r_k index behaves similar to the one in real graphs, reproducing the fragility evolution pattern; for the very first small graph snapshots, r_k is close to one indicating absence of robustness, while as the size of the graph increases, r_k decreases smoothly leading to snapshots with improved robustness index. As we can also observe from Fig. 11 (b) for these two values of r_k, the diameter of the graph initially increases for a few time points while the size of the graph is relatively small; however, after a specific point the diameter starts decreasing and almost stabilizes for the rest points. Notice that, the change point roughly corresponds to the change point of the r_k index (see also Section 6). For the last two values of the forward burning probability p_f, i.e., $p_f = 0.38$ and

$p_f = 0.4$, the FF model still approximates the evolution of the r_k index, but in a more "abrupt" way, where graphs tend to become extremely robust very quickly. Additionally, the evolution of diameter of these graphs is slightly different from the previous case ($p_f = 0.34$ and $p_f = 0.36$), where except from the smaller value, the shrinkage is more evident (especially for $p_f = 0.4$). To conclude, for values of r_k in the range $[0.34, 0.36]$ we observed that the FF model is able to reproduce the fragility evolution pattern. Note that, for a similar range of values, the authors of (Leskovec et al, 2009) observed that the FF model is able to reproduce the community structure of real graphs, as captured by the NCP plot (see also Paragraph 5.3).

8. Discussion and Conclusions

In this paper we studied the problem of estimating the robustness of social graphs, using the notion of expansion properties. Although our work focuses on social networks, the proposed robustness metric r_k can also be applied to other types of graphs from different disciplines (e.g., biological networks). The main contributions of this work are with respect to the following points:

- *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 that, relying on the spectral properties of real-world graphs, we can efficiently and effectively compute this measure, making it scalable for million-node graphs.
- *Application on real graphs*: 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. These two patterns give us further insights about the structure of large scale social graphs, and in particular the common regularities observed in most of them propose new structural properties (i.e., patterns) of real graphs.
- *Abnormality detection:* We showed how the observed patterns related to the r_k index can be used to detect anomalies in time-evolving graphs.
- *Robustness properties of graph generating models*: We studied several well-known graph generating models and we examined their ability to reproduce the observed robustness properties. Our results indicate that the Forest Fire model appears to produce graphs with robustness properties close to the observed ones, under appropriate settings of its parameters.

In addition to the above points, the proposed robustness measure can have further practical applications, and one of them concerns the community detection problem. In particular, the r_k index can give a fast estimation of the existence of "good" quality cuts in the network, and this knowledge can be further utilized by the community detection algorithm.

As we have already discussed, the proposed robustness estimation method is closely related to the existence of communities in real graphs. As future work, it would be interesting to examine how the r_k index is affected under the removal of nodes based on their degree (i.e., deletion of high degree nodes or randomly selected ones). This will help us to further study and understand the relationship of the proposed metric with the robustness assessment techniques from the area of network science (e.g., the work by Albert et al. (Albert et al, 2000)). Further-

more, another future research direction could be the extension of the method to the MapReduce framework for studying the robustness of billion-node graphs.

Acknowledgements. Fragkiskos D. Malliaros is a recipient of the Google Europe Fellowship in Graph Mining, and this research is supported in part by this Google Fellowship. Vasileios Megalooikonomou is partially supported by the ARMOR Project (FP7-ICT-2011-5.1 – 287720) which is co-funded by the European Commission under the Seventh Framework Programme and by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the NSRF - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund. Christos Faloutsos is supported by the National Science Foundation under Grants No. IIS-1217559 CNS-1314632, by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and under Contract Number W911NF-11-C-0088, by an IBM Faculty Award and a Google Focused Research Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

DBLP Bibliography Server (2006) http://dblp.uni-trier.de/xml/

KDD Cup (2004) http://www.cs.cornell.edu/projects/kddcup/

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

L. Akoglu, M. McGlohon, and C. Faloutsos (2010) OddBall: Spotting Anomalies in Weighted Graphs. In *PAKDD*, pp 410-421

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

R. Albert, H. Jeong, and A.-L. Barabasi (2000) Error and attack tolerance of complex networks. Nature, 406(6794):378382

A. Anagnostopoulos, G. Brova, and E. Terzi (2011) Peer and Authority Pressure in Information-Propagation Models. In *PKDD*, pp 76-91

R.A. Baeza-Yates and B. Ribeiro-Neto (1999) Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc.

A.-L. Barabási and R. Albert (1999) Emergence of Scaling in Random Networks. Science, 286(5439):509-512

B. Bollobás and O. Riordan (2003) Robustness and Vulnerability of Scale-Free Random Graphs. Internet Mathematics, 1(1):1-35

D.S. Callaway, M. E. J. Newman, S.H. Strogatz, and D.J. Watts (2000) Network Robustness and Fragility: Percolation on Random Graphs. Physical Review Letters, 80(25):5468-5471

D. Chakrabarti and C. Faloutsos (2012) Graph Mining: Laws, Tools, and Case Studies. Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan & Claypool Publishers

V. Chandola, A. Banerjee, and V. Kumar (2009) Anomaly detection: A survey. ACM Comput. Surv., 41(3):1-58

F.R.K. Chung (1997) Spectral Graph Theory. CBMS, Regional Conference Series in Mathematics, No. 92, AMS

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

R. Cohen and S. Havlin (2010) Complex Networks: Structure, Robustness and Function. Cambridge University Press

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

E. Estrada and J.A. Rodríguez-Velázquez (2005) Subgraph centrality in complex networks. Phys. Rev. E, 71(5):056103

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

E. Estrada (2006) Network robustness to targeted attacks. The interplay of expansibility and degree distribution. Eur. Phys. J. B 52:563-574
M. Faloutsos, P. Faloutsos, and C. Faloutsos (1999) On power-law relationships of the Internet topology. In *SIGCOMM*, pp 251-262
H. Fei and J. Huan (2008) Structure Feature Selection for Graph Classification. In *CIKM*, pp 991-1000
S. Fortunato (2010) Community detection in graphs. Physics Reports 486(3–5):75–174
G.H. Golub and C.F. Van Loan (1996) Matrix computations (3rd ed.). Johns Hopkins University Press
S. Hoory, N. Linial, and A. Wigderson (2006) Expander graphs and their applications. Bull. Amer. Math. Soc., 43:439-561
N. Jin and W. Wang (2011) LTS: Discriminative Subgraph Mining by Learning from Search History. In *ICDE*, pp 207-218
R. Kumar, J. Novak, and A. Tomkins (2006) Structure and evolution of online social networks. In *KDD*, pp 611-617
K. Lefevre and E. Terzi (2010) GraSS: Graph Structure Summarization. In *SDM*, pp 454-465
J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani Kronecker Graphs: An Approach to Modeling Networks Journal of Machine Learning Research, 11:985-1042
J. Leskovec, D. Huttenlocher, and J. Kleinberg (2010) Predicting Positive and Negative Links in Online Social Networks. In *WWW*, pp 641-650
J. Leskovec, J. Kleinberg, and C. Faloutsos (2005) Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD*, pp 177-187
J. Leskovec, J. Kleinberg, and C. Faloutsos (2007) Graph Evolution: Densification and Shrinking Diameters. ACM Trans. Knowl. Discov. Data 1(1)
J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney (2009) Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1):29-123
A.S. Maiya and T.Y. Berger-Wolf (2010) Expansion and search in networks. In *CIKM*, pp 239-248
F.D. Malliaros, V. Megalooikonomou, and C. Faloutsos (2012) Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly. In *SDM*, pp 942-953
F.D. Malliaros and Michalis Vazirgiannis (2013) Clustering and community detection in directed networks: A survey. Physics Reports 533(4):95–142
H. Maserrat and J. Pei (2010) Neighbor query friendly compression of social networks. In *KDD*, pp 533-542
M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen (2011) Sparsification of influence networks. In *KDD*, pp 529-537
M. McGlohon, L. Akoglu, and C. Faloutsos (2008) Weighted graphs and disconnected components: patterns and a generator. In *KDD*, pp 524-532
M. Mihail, C. Papadimitriou, and A. Saberi (2011) On certain connectivity properties of the Internet topology. In *FOCS*, pp 28-35
A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee (2007) Measurement and Analysis of Online Social Networks. In *IMC*, pp 29-42
B. Mohar (1989) Isoperimetric Number of Graphs. J. Comb. Theor. B 47(3):274-291
M.E.J. Newman (2003) The structure and function of complex networks. SIAM Review, 45:167-256
M.E.J. Newman and J. Park (2003) Why social networks are different from other types of networks. Phys. Rev. E 68:036122
M.E.J. Newman (2006) Finding community structure in networks using the eigenvector of matrices. Phys. Rev. E, 74(3):036104
M.E.J. Newman (2006) Modularity and community structure in networks. PNAS, 103(23):8577-8582
L. Page, S. Brin, R. Motwani and T. Winograd (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab
M. Richardson, R. Agrawal, and P. Domingos (2003) Trust Management for the Semantic Web. In *ISWC*, pp 351-368
A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B.Y. Zhao (2010) Measurement-calibrated graph models for social network experiments. In *WWW*, pp 861-870
V. Satuluri and S. Parthasarathy (2009) Scalable graph clustering using stochastic flows: applications to community. discovery. In *KDD*, pp 737–746

C. Seshadhri, Ali Pinar, and Tamara G. Kolda (2013) An In-Depth Analysis of Stochastic Kronecker Graphs. J ACM, 60(2):13:1–13:32
H. Toivonen, F. Zhou, A. Hartikainen and A. Hinkka (2011) Compression of weighted graphs. In *KDD*, pp 965-973
C.E. Tsourakakis (2008) Fast Counting of Triangles in Large Real Networks without Counting: Algorithms and Laws. In *ICDM*, pp 608-617
C.E. Tsourakakis (2011) Counting triangles in real-world networks using projections. Knowl Inf Syst, 26:501-520
B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi (2009) On the Evolution of User Interaction in Facebook. In *WOSN*, pp 37-42
D.J. Watts and S.H. Strogatz (1998) Collective dynamics of 'small-world' networks. Nature, 393(684):440-442
X. Yan and J. Han (2002) gSpan: Graph-Based Substructure Pattern Mining. In *ICDM*, pp 721-724

Appendix

In this Appendix, we provide a more detailed description of how the property of large spectral gap along with the subgraph centrality measure, lead to the measure $\xi(G)$ (Estrada, 2006) as presented in Section 3. First of all, the subgraph centrality measure is defined as (Estrada and Rodríguez-Velázquez, 2005)

$$SC(i) = \sum_{\ell=0}^{\infty} \frac{A_{ii}^{\ell}}{\ell!}, \ \forall i \in V, \tag{5}$$

where the diagonal entry A_{ii} of the matrix $\mathbf{A}^{\ell}$ contains the number of closed walks of length ℓ that begin and end at the same node i. Focusing on unipartite graphs and keeping only the odd length closed walks[4] in order to avoid cycles in acyclic graphs, the SC can be expressed as

$$SC(i) = u_{i1}^2 \sinh(\lambda_1) + \sum_{j=2}^{|V|} u_{ij}^2 \sinh(\lambda_j). \tag{6}$$

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. (6) could be written as

$$SC(i) \approx u_{i1}^2 \sinh(\lambda_1), \ \forall i \in V. \tag{7}$$

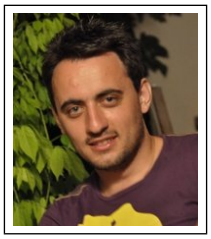
This means that for graphs with high robustness, the principal eigenvector u_{i1} will be related to $SC(i)$ as

$$u_{i1} \propto \sinh^{-1/2}(\lambda_1) \ SC(i)^{1/2}. \tag{8}$$

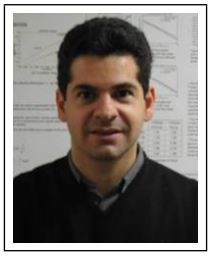
[4] The bipartite graphs do not have odd length closed walks and thus the SC is computed based on the even length closed walks. This happens replacing the $\sinh(\cdot)$ function with the $\cosh(\cdot)$ (Estrada and Rodríguez-Velázquez, 2005). But then the SC for the bipartite graphs cannot be efficiently approximated using similar ideas with the proposed NSC_k (Section 4), because of the fact that the $\cosh(\cdot)$ is an even function. However, our approach for bipartite graphs (Section 4, Proposition 4.1) overcomes this bottleneck and can be efficiently computed for large scale graphs.

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).

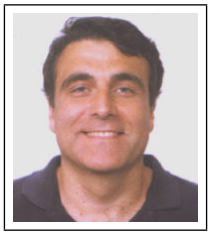
Author Biographies



Fragkiskos D. Malliaros is currently a Ph.D. candidate in the Computer Science Laboratory at École Polytechnique in France, working under the supervision of Prof. Michalis Vazirgiannis. He received his Diploma and his M.Sc. degree from the Computer Engineering and Informatics Department of the University of Patras, Greece in 2009 and 2011 respectively. He is the recipient of the 2012 Google European Doctoral Fellowship in Graph Mining. During the summer of 2014, he was a research intern at the Palo Alto Research Center (PARC). He has also published six referred articles in international data mining venues and journals. His research interests span the broad areas of data mining, algorithmic data analysis and data management, with current focus on mining and analysis of real-world graphs.



Vasileios Megalooikonomou received a BSc in computer engineering and informatics from the University of Patras, Greece in 1991, and a M.S. and Ph.D. in computer science from the University of Maryland, Baltimore County, USA, in 1995 and 1997, respectively. He has been on the faculty of Johns Hopkins University, Dartmouth College, Temple University and University of Patras, Greece. He has co-authored over 150 refereed articles in journals and conference proceedings and two book chapters. His main research interests include biomedical informatics, data mining, data compression, pattern recognition, and multimedia database systems. He is a member of the ACM, IEEE, SIAM, and SPIE. In 2003 he received a CAREER award from the National Science Foundation for developing data mining methods for extracting patterns from medical image databases. He regularly serves as a program committee member or referee on several premier conferences and journals in his areas of research.



Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010), nineteen "best paper" awards (including two "test of time" awards), and four teaching awards. He is an ACM Fellow, he has served as a member of the executive committee of SIGKDD; he has published over 200 refereed articles, 11 book chapters and one monograph. He holds six patents and he has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bio-informatics data.

Correspondence and offprint requests to: Fragkiskos D. Malliaros. Laboratoire d'Informatique (LIX), Bâtiment Alan Turing, 1 rue Honoré d'Estienne d'Orves, Campus de l'École Polytechnique, 91120 Palaiseau, France. Tel.: +33 0177578045. Email: fmalliaros@lix.polytechnique.fr