How Robust is the Core of a Network?

Abhijin Adiga[1] and Anil Kumar S. Vullikanti[1,2]

[1] Virginia Bioinformatics Institute, Virginia Tech
[2] Department of Computer Science, Virginia Tech
abhijin@vbi.vt.edu, vsakumar@vt.edu

Abstract. The k-core is commonly used as a measure of importance and well connectedness for nodes in diverse applications in social networks and bioinformatics. Since network data is commonly noisy and incomplete, a fundamental issue is to understand how robust the core decomposition is to noise. Further, in many settings, such as online social media networks, usually only a sample of the network is available. Therefore, a related question is: How robust is the top core set under such sampling?
We find that, in general, the top core is quite sensitive to both noise and sampling; we quantify this in terms of the Jaccard similarity of the set of top core nodes between the original and perturbed/sampled graphs. Most importantly, we find that the overlap with the top core set varies *non-monotonically* with the extent of perturbations/sampling. We explain some of these empirical observations by rigorous analysis in simple network models. Our work has important implications for the use of the core decomposition and nodes in the top cores in network analysis applications, and suggests the need for a more careful characterization of the missing data and sensitivity to it.

1 Introduction

The k-core $C_k(G)$ of an undirected graph $G = (V, E)$ is defined as the maximal subgraph in which each node has degree at least k; the core number of a node is the largest k such that it belongs to the k-core (i.e., $v \in C_k(G)$). The set $S_k(G) = C_k(G) \setminus C_{k+1}(G)$, consisting of nodes with core-number k, is referred to as the k-shell; the core decomposition (i.e., the partitioning into shells) can be computed efficiently and combines local as well as global aspects of the network structure. This makes it a very popular measure (along with other graph properties, e.g., degree distribution and clustering coefficient) in a wide variety of applications, such as: the autonomous system level graph of the Internet [6,3], bioinformatics [18,26], social networks and epidemiology [20,17]; some of the key properties that have been identified include: the well-connectedness of the nodes with high core number and their significance in controlling cascades.

In most applications however, the networks are inferred by indirect measurements, e.g.: (i) the Internet router/AS level graphs constructed using traceroutes, e.g., [12], (ii) biological networks, which are inferred by experimental correlations, e.g., [18,26], (iii) networks based on Twitter data (related to which there is a

growing body of research, e.g., [17,4,16]), in which a limited 1% sample can be constructed by the APIs.[3] Therefore, networks studied in these applications are inherently noisy and incomplete; this raises a fundamental issue in the use of any graph property $\mathcal{P}(G)$ for graph G: How does the property, and conclusions based on it get affected by the uncertainty in G? Is there a smooth transition in the property with the uncertainty,[4] and is it possible to quantify the error in the observed measurement? An example of such an issue is the nature of degree distributions of the Internet router graph and its vulnerability: several papers, e.g., [12] observed that these are power laws. Achlioptas et al. [1] showed that there are significant sampling biases in the way traceroutes (which are used to infer the network) work; for a broad class of networks, they prove that such inference methods might incorrectly infer a power-law distribution (even when the underlying network is not).

Our work is motivated by these considerations of the sensitivity to noise and the adequacy of sampling. Specifically, we study how results about the core decomposition and top cores in the network, e.g., [6,3,18,26,20,17], are affected by the uncertainty, noise and small samples (as in the case of online social media networks). Such questions have been studied in the statistical physics literature, e.g., [10], who show that there is a threshold probability for random node deletions in infinite networks, above which the k-core disappears; it is not clear how relevant such results are to real world networks, which are finite and do not satisfy the symmetries needed in such results. Hamelin et al. [3] report robustness of their observations related to the shell structure in the Internet router graph, for specific sampling biases related to traceroute methods. We are not aware of any other empirical or analytical work on the sensitivity of the core decomposition.

Since there is very limited understanding of how noise should be modeled, we consider three different stochastic edge perturbation models, which are specified by how a pair u, v of nodes is picked: (i) uniformly at random (`ERP`, for Erdős-Rényi perturbations), (ii) in a biased manner, e.g., based on the degrees of u, v (`CLP`, for Chung-Lu or degree assortative perturbations), and (iii) by running a missing link prediction algorithm, such as [8] (`LPP`, for link prediction based perturbations); see Section 3 for complete definitions. We also study a model of stochastic node deletions. Let α denote the fraction of nodes/edges perturbed; typically we are interested in "small" α.

A complementary aspect (particularly relevant in the context of sampled data from social media such as Twitter) is the effect of sampling. We consider edge/node sampling with probability p (i.e., corresponding to deletion with probability $1 - p$). We study the following question: can the properties of the core structure be identified by small edge/node samples, i.e., corresponding to small p? In our discussion below, we use G' to denote the graph resulting from per-

[3] Larger samples, e.g., 10% can be obtained form Twitter's commercial partners for a large fee.

[4] As observed in the case of centrality measures by [5], who claim that there is a gradual decrease in the accuracy of the centrality scores.

turbation/sampling of a starting graph G. Let $k_{max}(G)$ denote the maximum core number in G. We study the Jaccard similarity, $\eta_j(G, G')$, between the set of nodes in the top j-cores in G and G'; we sometimes informally refer to η_j as the "similarity" between the top cores. Our main results are summarized below.

1. Sensitivity to Noise. We consider perturbations with α ranging from less than 1% to 10%. We find that η_j and the shell structure shows high sensitivity to edge/node perturbations; however, the precise effects are very network and noise model specific. Further, η_1 is quite sensitive in the `CLP` model for many networks: perturbation with $\alpha < 5\%$ can alter η_1 by more than 20% in some networks. More importantly, we find that in a large fraction of the networks, η_j exhibits a *non-monotone* behavior as a function of α. This can be a serious issue in some applications where the core structure is used, and needs to be examined critically. The sensitivity decreases as we increase j, but η_j varies non-monotonically with j as well. In contrast, the top cores seem quite stable to perturbations in the `ERP` model, which primarily affects the shell size distribution in some networks. The `LPP` model seems to affect both the low and high cores. Further, node perturbations (modeled as random deletions) seem to have a much higher impact than edge perturbations, in general. It is intriguing that co-authorship and citation networks seem to be generally much more stable compared to influence and infrastructure networks. Further, we observe that sudden changes in the similarity index are almost always accompanied with increase in k_{max}.

This motivates the CorePerturbation problem: given a graph G and a parameter k, what is the probability that a k-core forms in G after perturbation, if it did not have a k-core initially. We prove that this problem is #P-hard, which suggests rigorous quantification of the variation in the top core even in such simple stochastic noise models is quite challenging. We attempt to further understand and explain the empirical observations analytically using simple mathematical models. We also prove that under some weak assumptions that usually hold in social networks, the low core numbers can be altered quite significantly in the `ERP` model.

2. Sensitivity to Sampling. We find most networks exhibit a high level of sensitivity to sampling, and η_j is a noisy and non-monotone function of p, especially when p is close to 1; there is higher level of sensitivity to node sampling than to edge sampling. For most of the networks we study, identifying a reasonably large fraction (say 80%) of the top core set requires a fairly high sampling rate p: higher than 0.6 in most networks, and as high as 0.8 in some. Specifically, in the case of a Twitter "mentions" graph (see Section 3.2 for details), we find that this entails a much higher level of sampling than what is supported by the public API. Further, biased sampling based on edge weights can improve the similarity index slightly. We analyze the effects of sampling to help explain some of these results. We show that the maximum core number in G_p scales with the sampling probability, and that non-monotonicity under sampling is an inherent aspect of the Erdős-Rényi model. We also find that the top core can be very fragile, and can change completely even for very low sampling rate.

Organization. We briefly discuss the related work in Section 2. We introduce the main definitions, and summarize our data sets in Section 3. We discuss the sensitivity to noise and the effects of sampling in Sections 4 and 6, respectively. In Section 5, we discuss the CorePerturbation problem, and conclude in Section 7. Because of limited space, we present many of the details in the full version [2].

2 Related Work

Noise and sampling biases in networks are well recognized as fundamental issues in applications of complex networks, and many different models have been studied for it. A common approach in social networks, e.g., [9,5], is to examine stochastic node and edge deletions. There is a large body of work on predicting missing links in complex networks (based on expected clustering and other structural properties), e.g., [8], which could also be used as a possible candidate set. Since there is no clear understanding of noise/perturbations, we study three different models from the literature in this paper.

We briefly discuss a few of the results on understanding the impact of uncertainty on network properties. The impact of sampling bias on the properties of the Internet router graph [1] was already mentioned earlier. There has also been a lot of work in understanding the sensitivity of centrality to noise, e.g., [9,5]; it has been found that the impact on the centrality is variable and network dependent, but the general finding in [5] is that the accuracy of centrality measures varies smoothly and predictably with the noise. Morstatter et al. [22] study the effects of limited sampling in social media data by analyzing the differences in statistical measures, such as hashtag frequencies, and network measures, such as centrality.

The work by Flaxman and Frieze [14,15] is among the very few rigorous results on the impact of perturbations on network parameters— they rigorously analyze the impact of `ERP` on the diameter and expansion of the graph. The issue of noise has motivated a number of sampling based algorithmic techniques which are "robust" to uncertainty, in the form of "property testing" algorithms, e.g., [25] and "smoothed analysis", e.g., [27].

Finally, we briefly discuss some of the work on the core decomposition in graphs. As mentioned earlier, the core number and the top core set has been used in a number of applications, e.g., [6,3,18,26,20,17], in which the shell structure and the top core sets have been found to give useful insights. Conditions for existence of the k-core, and determining its size have been rigorously studied in different random graph models, e.g., [24,13]; the main result is that there is a sharp threshold for the sudden emergence of the k-core in these models. This has also been analyzed in the statistical physics literature, e.g., [10]; these papers also study the impact of node deletions on the core size in infinite graphs, and show a characterization in terms of the second moment of the degree distribution.

3 Definitions and Notations

The k-core $C_k(G)$ of an undirected graph $G = (V, E)$ is defined as the maximal subgraph of nodes in which each node has degree at least k; the core-number of a node v is the largest k such that $v \in C_k(G)$. The set $S_k(G) = C_k(G) \setminus C_{k+1}(G)$ is referred to as the kth-shell of G; we omit G, and refer to it by S_k, when the graph is clear from the context. The set $C_k(G)$, if it exists, can be obtained by repeatedly removing vertices of degree less than k until no further removal is possible. The maximum k such that $C_k(G) \neq \phi$ will be denoted by $k_{\max}(G)$; we use just $k_{\max}$, when there is no ambiguity. The core decomposition of a graph G corresponds to the partition $S_0, S_1, \ldots, S_{k_{\max}}$ of V. Let $s_i = |S_i|$. We use $\beta(G) = \langle s_1, s_2, \ldots, s_{k_{\max}} \rangle$ to denote the vector of shell size distribution in G. The *Jaccard index* is a measure of similarity between two sets and is defined as follows: For sets A and B, $JI(A, B) = \frac{|A \cap B|}{|A \cup B|}$. In our empirical analysis of networks we compare the top j cores of the unperturbed and the perturbed graphs using the Jaccard index. For this purpose we introduce the notation $\eta_j(G, G') := JI\left(\cup_{i \geq k_{\max}-j+1} C_i(G), \cup_{i \geq k_{\max}-j+1} C_i(G')\right)$. The *variation distance* between the core-number distributions of two graphs G and G' on the same vertex set V is defined as, $\delta(G, G') = \frac{1}{2|V|} \sum_i |s_i(G) - s_i(G')|$. We say that an event holds **whp** (with high probability) if it holds with probability tending to 1 as $n \to \infty$.

3.1 Noise Models

Since there is no clear understanding of how uncertainty/noise should be modeled, we introduce a generalized noise model for edge perturbations which captures most models in literature, and also enables us to control separately the extent of addition and deletion. Let G be the unperturbed graph. Let $\mathbb{G} = \mathbb{G}(n)$ denote a random graph model on n nodes which is specified by the probability $P_{\mathbb{G}}((u, v))$ of choosing the edge (u, v). We define a noise model $\mathcal{N}(G, \mathbb{G}, \epsilon_a, \epsilon_d)$ based on $\mathbb{G}$ as a random graph model where the edge probability between a pair u, v is given by

$$P_{\mathcal{N}}((u, v)) = \begin{cases} \epsilon_a P_{\mathbb{G}}((u, v)), \text{ if } (u, v) \notin E_G, \\ \epsilon_d P_{\mathbb{G}}((u, v)), \text{ if } (u, v) \in E_G, \end{cases} \tag{1}$$

where ϵ_a and ϵ_d denote the edge addition and deletion probabilities, respectively. The perturbed graph $G' = G \oplus R$ is obtained by XORing G with $R \in \mathcal{N}(G, \mathbb{G}, \epsilon_a, \epsilon_d)$, a sample from the noise model, i.e., if $(u, v) \in E_G$, then it is deleted with probability $\epsilon_d P_{\mathbb{G}}((u, v))$, but if $(u, v) \notin E_G$, (u, v) is added with probability $\epsilon_a P_{\mathbb{G}}((u, v))$. Depending on how we specify $P_{\mathbb{G}}$ and the parameters ϵ_a, ϵ_d, we get different models; we consider three specific models below.

Uniform Perturbation (`ERP`): In this model we set $\mathbb{G} = \mathcal{G}(n, 1/n)$, the Erdős-Rényi random graph model where each edge is chosen with probability $1/n$ independently, i.e., $P_{\mathbb{G}}((u, v)) = 1/n$. We use the following notation for this model: $\texttt{ERP}(G, \epsilon_a, \epsilon_d) = \mathcal{N}(G, \mathcal{G}(n, 1/n), \epsilon_a, \epsilon_d)$. For example, $\texttt{ERP}(G, \epsilon, \epsilon)$ corresponds to

adding an edge or removing an existing edge independently with probability ϵ/n, while $\texttt{ERP}(G, \epsilon, 0)$ corresponds to only adding edges. This is the simplest model, and has been studied in social network applications, e.g., [9,5].

Degree Assortative Perturbation (CLP): In this model, $\mathbb{G}$ corresponds to the Chung-Lu random graph model [7] for graphs with a given expected degree sequence. Each node u is associated with a weight w_u (which we take to be its degree), and edge is chosen independently with probability proportional to the product of the weights of its endpoints, i.e., $P_{\mathbb{G}}((u,v)) \propto w_u \cdot w_v = d(u) \cdot d(v)$. This model selects edges in a biased manner, and might be suitable in applications dealing with assortative graphs with correlations between degrees of the end points of edges, which has been observed in a number of networks, e.g., [23].

Link Prediction Based Model (LPP): Instead of the purely stochastic ERP and CLP models, we use the results of a missing link prediction algorithm to determine which edges to perturb. Here, we use the algorithm of Clauset, et al. [8], which has been used quite extensively in the social network literature; further, since it uses a hierarchical random graph model, it can be viewed as an instance of our generalized noise model. This model is based on the assumption that many real-life networks have a hierarchical structure, which can be represented by a binary tree with n leaves corresponding to the node (referred to as a "dendrogram"). Given such a dendrogram D, each internal node r is associated with a probability $p_r = \frac{E_r}{L_r R_r}$, where L_r and R_r are the number of leaves in the left and right subtrees of r respectively and E_r is the number of edges between L_r and R_r in G. The likelihood of D is defined as: $\mathcal{L}(D) = \Pi_r p_r^{E_r}(1-p_r)^{L_r R_r - E_r}$. The algorithm of [8] specifies the probability, $P_D((u,v))$, of an edge between two vertices u, v, to be the value p_r, where r is the lowest common ancestor of u and v in D.

In the ERP model, the expected number of perturbed edges is $\approx n\epsilon/2$. For the purpose of fair comparison of noise models, we have normalized the weights of vertices in the CLP model such that the expected number of perturbed edges is again $\approx n\epsilon/2$. We use G_ϵ to denote the perturbed network. In the LPP model, we add edges as prescribed [8]; $n\epsilon/2$ edges are added in the decreasing order of their associated probabilities.

Additions vs Deletions: We find that, due to sparsity of the networks considered, perturbations involving edge additions/deletions do not alter the results by much, compared to perturbations involving just edge additions. Hence, unless explicitly specified, we only consider addition of edges. Also, henceforth, whenever we use the truncated notations ERP and CLP, we refer to $\texttt{ERP}(G, \epsilon, 0)$ and $\texttt{CLP}(G, \epsilon, 0)$, respectively.

Noise could also manifest in terms of missing nodes. We study a model of random node deletions with probability $1-p$ (which corresponds to retaining nodes with probability p); we study the effect of this in the form of sampling in Section 6, instead of perturbations.

3.2 Data

In order to make our results as robust as possible, we analyze over 25 different real (from [21]) and random networks. We also used a Twitter mentions graph, constructed in the following manner: we consider a set of about 9 million tweets (corresponding to a 10% sample, obtained from a commercial source), and construct a graph on the Twitter users, in which an edge (u, v) denotes a mention of user v by user u (in the form of an "@v" in the tweet) or the other way around; this graph has over 2 million nodes and about 4.6 million edges. We then considered subgraphs constructed by sampling edges with probability $p \in [0.1, \ldots, 0.99]$; for $p \in [0.1, \ldots, 0.8]$, we use increments of 0.1, but for $p \in [0.8, 0.99]$, we use increments of 0.01, in order to increase the resolution. Finally, we also consider random graph models with Poisson and scale-free degree distributions. Table 1 contains a summary of the graphs analyzed.

Table 1. Real-world and synthetic graphs used in our experiments and their properties.

Class	Network	N	E	$k_{\max}$	$\lvert C_{k_{\max}}(G)\rvert$
Autonomous Systems	As20000102	6474	12572	12	21
	Oregon1010331	10670	22002	17	32
	Oregon2010331	10900	31180	31	78
Co-authorship	Astroph	17903	196972	56	57
	Condmat	21363	91286	25	26
	Grqc	4158	13422	43	44
	Hepph	11204	117619	238	239
	Hepth	8638	24806	31	32
Citation	HepPh	34546	420877	30	40
	HepTh	27770	352285	37	52
Communication	Email-EuAll	265214	364481	37	292
	Email-Enron	33696	180811	43	275
Social	Epinion	75877	405739	67	486
	Slashdot0811	77360	469180	54	129
	Soc-Slashdot0902	82168	504230	55	134
	Twitter	22405	59898	20	177
	Wiki-Vote	7066	100736	53	336
	Twitter "mentions"	2616396	4677321	19	210
Internet peer-to-peer	Gnutella04	10876	39994	7	365
	Gnutella24	26518	65369	5	7480
Synthetic graphs	Regular ($d = 20$)	10000	100000	20	10000

4 Sensitivity of the Core Decomposition to Noise

We now study the effect of node/edge perturbations on the similarity index $\eta_j(G, G')$, and the changes in the shell size distribution $\beta(G)$ in terms of the variation distance, $\delta(G, G')$ (see Section 3 for definitions). We study these quantities on the networks mentioned in Section 3.2 and for the perturbation models discussed in Section 3.1. For the `ERP` and `CLP` models, we compute 100 to 1000 instances, for each choice of ϵ, over which $\eta_j(G, G')$ and $\delta(G, G')$ are averaged. The methodology for the `LPP` model is discussed later.

4.1 Sensitivity of the Top Cores

1. Sensitivity of the Top Core in the `CLP` *Model:* Figure 1 shows the variation in $\eta_1(G, G')$ for different networks in this model. The figure shows the variation with both ϵ and α (the fraction of edges added), the latter to account for the difference in the graph sizes. The most striking observation is the high sensitivity of η_1 and its highly non-monotonic variation in a large fraction of the networks. The specific points where significant jumps in η_1 happen correspond to the points where k_{max} changes in many cases, as shown in Figure 1(c). The specific behavior is highly variable and network dependent. For example, we note that while the top cores in collaboration and citation networks are, in general, highly resilient to perturbation, most social and peer to peer networks show great variation.

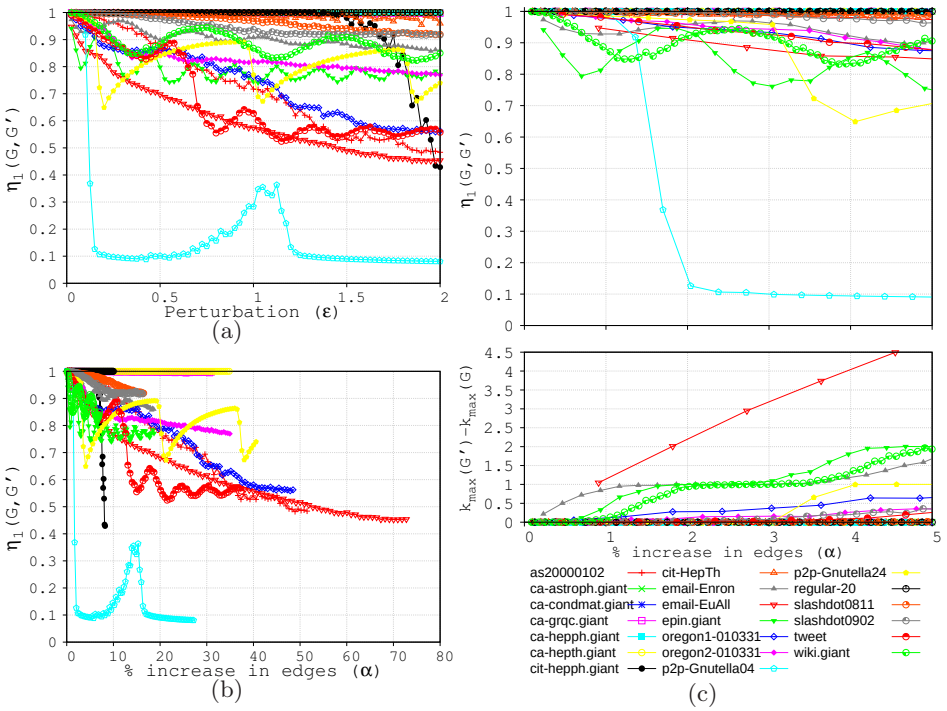

Fig. 1. Top core comparison for various networks under degree-weighted edge perturbation `CLP`: Here, (c) is a zoomed plot of (b). This is complemented by a plot of $k_{\max}(G') - k_{\max}(G)$ to depict the transition to a higher core and its effect on $\eta_1(G, G')$.

2. Sensitivity of the top core in the `ERP` *Model:* In contrast to the `CLP` model, we find that top cores are much more stable in the `ERP` model. The main reason for this stability is the fact that almost all networks considered here have very small fraction of nodes in the top core(s) (as shown in Figure 7 in the full version [2]), so that most of the edges in the `ERP` model are added to low core nodes

3. Explaining the Differences Between the CLP and ERP Models: We note that in the CLP model, the higher the degree of a node in the unperturbed graph, the greater is the number of edges incident with it after perturbation. This polarizing nature of the model needs to be taken into account to infer and quantify the stability of the top core. Figure 6 in the full version [2] shows scatter plots of core number vs. degree for some selected graphs. Even though it gives some idea about the behavior of the top core, we find it highly non-trivial to quantify the stability in any way. Later, in Section 5 we will be considering a theoretical formulation of this problem and showing that such a quantification of stability is in general hard.

4. Sensitivity of the Top 5 Cores: We extend our empirical analysis to $\eta_j(G, G')$ for $j = 1, \ldots, 5$ in Figure 2. Note that the non-monotonic behavior is mitigated in these plots, but η_j varies non-monotonically with j. However, as j is increased, the size of C_j can become very large, thus diminishing the main utility of the top cores in most applications.

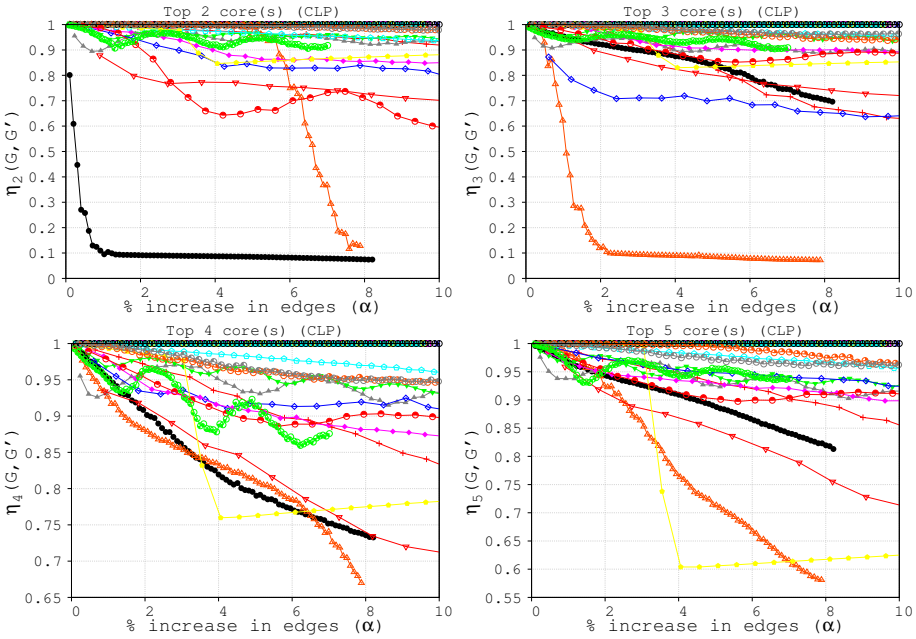

Fig. 2. Top 2–5 cores comparison ($\eta_j(G, G')$, $j = 2, \ldots, 5$) with respect to % increase in edges (α). The legends are the same as in Figure 1.

5. Sensitivity in the LPP Model: We considered the stability of the top cores in the LPP model by applying the link prediction algorithm given in [8]. We first generated the list of likelihood probabilities for each possible edge. For this purpose, we used the implementation of [11]. The edges were then added in the descending order of their probabilities. As shown in Figure 3(a) for a subset of graphs, the variation in η_1 is very network specific, and hard to characterize.

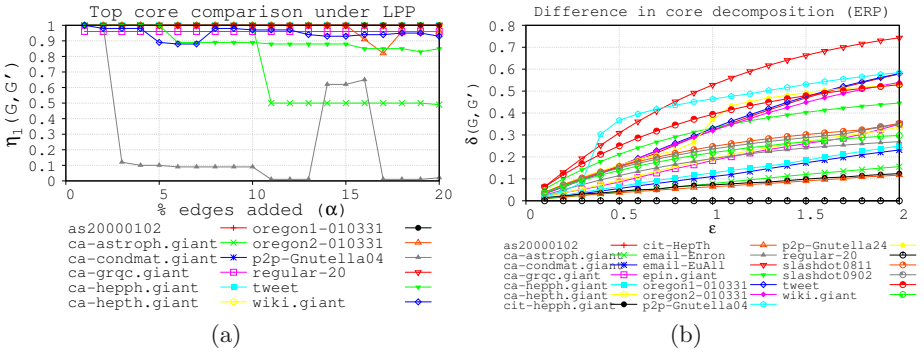

Fig. 3. Core stability with respect to LPP and ERP.

4.2 Sensitivity of the Shell Size Distribution and the Low Cores

To study the effect of perturbation on the core decomposition of a network, we consider the variation distance $\delta(G, G')$ (defined in Section 3). The results are in Figure 3(b). As discussed above, the ERP model has greater impact on the overall core-structure compared to the CLP model; in the ERP model, changes happen to lower core structure which contain most of the nodes and hence leads to large variation distance. We observe no significant change in $\delta(G, G')$ under the CLP model, as is evident from Figure 7 in [2].

We attempt to explain some of the observations about the changes in the core structure analytically. First, we consider the impact of perturbations on the 2-core in any graph in the ERP model, and prove that for any constant $\epsilon > 0$, the 2-core always becomes of size $\Theta(n)$, which is consistent with the results in Figure 7 in [2]. Our results are similar in spirit to the work of [14]. This only explains the changes in the lowest core, and in order to extend it further, we examine a quantity motivated by the "corona" [10], which corresponds to nodes which need few edges to the higher cores in order to alter the core number. We find that there is a large fraction of nodes of this kind in many networks, which might help in characterizing the stability of the shell structure. This is discussed in [2].

Theorem 1. *Let G be any connected graph with n vertices and let $G_p = G \oplus R$ where, $R \sim \mathcal{G}(n, \epsilon/n)$ and ϵ is a constant. Then,* **whp** *G_p has a 2-core of size $\Theta(n)$.*

Proof. (Sketch) Consider a spanning tree T of G. We show that $T \oplus R$ itself has a $\Theta(n)$ sized 2-core. Let T^- denote the subgraph obtained by removing the edges common to T and R. Suppose e_d is the number of edges removed from T. Since each edge of T can be removed with probability ϵ/n, it can be verified that **whp** $O(\log n)$ edges are removed from T, so that T^- has $O(\log n)$ components **whp**.

Let I be a maximum independent set of T^-. We consider the graph induced by I in R and consider the edges of $R[I]$ (not belonging to T) one at a time.

Let $e_i = (u, v)$ be the ith edge of $R[I]$ added to T^-. Let $T_i^- = T_{i-1}^- + e_i$ with $T_0^- = T^-$. If u and v belong to the same component of T_{i-1}^-, then, there is a path P in T_{i-1}^- with end points u and v and therefore, u and v both belong to the 2-core in $T \oplus R$. However, when u and v belong to different components, this does not happen. However, note that each time this happens, T_i^- has one less component compared to T_{i-1}^-. Since T_0^- has $O(\log n)$ components, there can be at most $O(\log n)$ such edges.

Consider vertices in $R[I]$ of degree at least 1. Note that $|I| = cn$ for some constant $c \geq 1/2$. It is easy to see that **whp** a constant fraction of these vertices in I have degree at least 1 in $R[I]$. Of these vertices, we will discard vertices which are end points of an edge e_i which is between two components of T_{i-1} for some i. From the previous discussion, there can be only $O(\log n)$ such edges. The rest of the vertices form a 2-core. Hence proved. □

5 The CorePerturbation Problem

From Section 4.1, it follows that the sudden and non-monotone changes in the similarity index correspond to an increase in the maximum core number. This motivates the CorePerturbation problem, which captures the probability of this change happening.

Definition 1. *The* CorePerturbation *problem (*$\mathrm{CP}(G, E_A, p, k)$*)*
Input: *A graph* $G(V, E)$*, an integer* $k \geq 4$*, edge probability* p *and a set of possible edges* E_A *(which are absent in* G*). Let* G_p *be the graph resulting from adding edges to* G *from* E_A *independently with probability* p*.*
Output: *Probability that* G_p *has a* k*-core.*

Theorem 2. $\mathrm{CP}(G, E_A, p, k)$ *is #P-complete.*

The proof of Theorem 2 is in the full version [2]. The result also holds when $k_{\max}(G) = k - 1$, which implies that even in a very simple noise model, quantifying the precise effects of changes in the top core is very challenging. When this probability is not too small (e.g., larger than $1/n^c$ for some constant $c > 0$), it can be shown that a polynomial number of Monte-Carlo samples can give good estimates (within a multiplicative factor of $1 \pm \delta$, with any desired confidence, where $\delta > 0$ is a parameter).

6 Sensitivity of the Core Decomposition to Sampling

We now address the issue of sampling and focus on $\eta_k(G, G_p)$, where G_p denotes a node/edge sampled graph with probability p— our goal is to understand to what extent the core structure (especially the nodes in the top cores) can be identified from sparsely sampled data. As in the case of noise (Section 4), we find η_k is quite sensitive to sampling, and varies non-monotonically for many graphs. We attempt to explain these results rigorously in the following manner:

(i) using the notion of edge density, we derive bounds on the maximum core in sampled graphs, which show that it scales with p, (ii) we analyze the sampling process in random graphs, and prove that the non-monotonicity in η_k is an inherent issue related to the core structure.

6.1 Variation in η_k

Figure 4(a) shows the variation in $\eta_1(G, G_p)$ for all networks, for an edge sampling probability $p \in [0.8, 1]$. We observe that η_1 is quite low in many networks; in order to identify at least 80% of the top core nodes (i.e., $\eta_1 \geq 0.8$), we need $p \geq 0.6$ in most networks. Figure 9 in the full version [2] shows additional results on the effect of edge sampling on η_k, for $k \in \{1, 2, 5, 10\}$. Like in the case of edge perturbations, we find η_k also exhibits non-monotonicity with respect to k for most networks. Further, it is interesting to note that the citation networks are very sensitive to sampling (and have η_1 below 0.6), but were found to be quite robust to edge perturbations (Section 4). However, collaboration networks seem to be robust to sampling as in the case of edge perturbations. We find that node sampling has a much higher impact than edge sampling; see Table 2 in the full version [2] for details. For instance, with $p = 0.95$, we observe that η_1 is below 0.9 for all but four of the networks, and is below 0.62 in three networks.

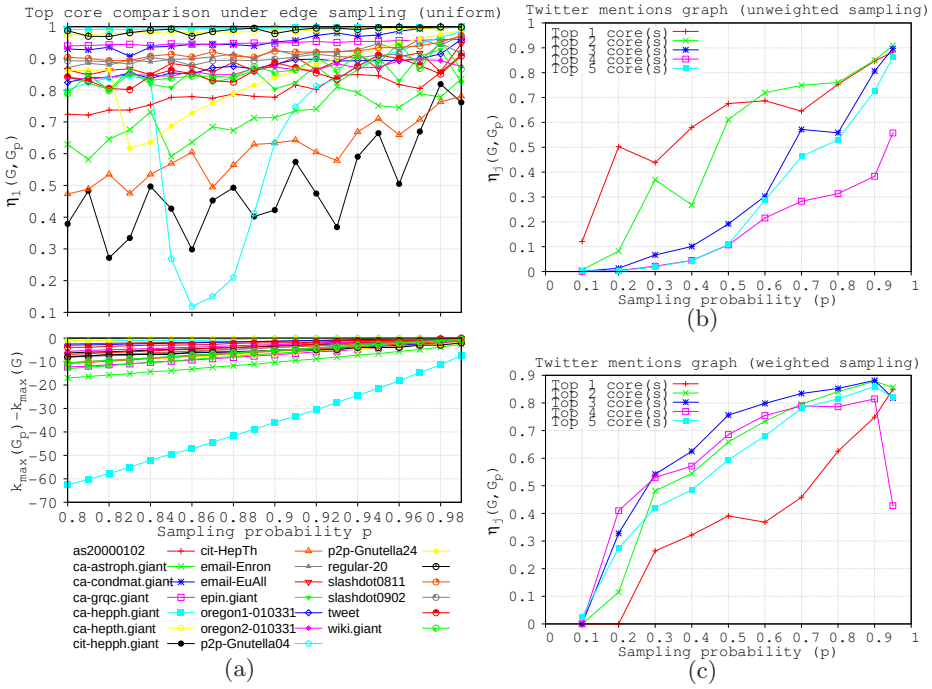

Fig. 4. Top core comparison for various networks under sampling edges.

Biased Sampling in Twitter Networks. Since sampling is an inherent aspect of the APIs provided by Twitter, we study its effects on the top cores. Our results for the Twitter mentions graph (see Section 3.2 for the details) are shown in Figures 4(b) and 4(c). The graph is weighted, in which the weight of an edge (u, v) corresponds to the number of mentions of u by v (or the converse). We observed that η_k is generally quite low, but is somewhat higher when edges are sampled with probability proportional to the edge weight (Figure 4(b)) instead of uniform (Figure 4(c)). Moreover, there is high non-monotonicity in both scenarios, suggesting that Twitter's public API is not adequate for identifying the core structure with high confidence (say 80% or more), and multiple calls to the API must be run to improve the accuracy. Table 3 in the Appendix of [2] gives additional details on the max core values in the sampled graphs.

Bounding the Max Core on Sampling. A first step towards understanding the effect of sampling is to determine $k_{\max}(G_p)$ in the sampled graph G_p. Table 3 in the full version [2] suggests that $k_{\max}$ scales with the sampling probability. This is examined in the following lemma, whose proof is discussed in the full version [2].

Lemma 1. *Consider a graph G such that $k_{\max}(G) \to \infty$ as $n \to \infty$. Let G_p denote the random subgraph of G obtained by retaining each edge of G with probability p, where p is a constant. Then, for any constant $\delta \in (0, 1)$, $k_{\max}(G_p) > (1 - \delta)k_{\max}(G)p/2$,* **whp**.

6.2 Core Structure in Random Graphs

In order to understand our empirical observations about the sensitivity of the core structure to noise and sampling, and especially the non-monotone behavior, we now study the effect of sampling in random graph models. We consider two random graph families: (a) the Erdős-Rényi random graphs and (b) Chung-Lu power law random graphs [7] with node weights picked from a power-law distribution (see Section 3 for a description of this model). Figure 5 shows the sensitivity of $\eta_k(G, G_p)$ to the sampling probability p. Figure 5(a) shows the results for a random graph from $\mathcal{G}(n, p)$ for $n = 10000$ and $p = 50/n$ and Figure 5(b) shows the results for a graph from the Chung-Lu model with power law exponent 2.5, $n = 10000$ and average degree 5. We observe non-monotone variation in η_k with p; this is more pronounced in the case of the Chung-Lu model, in which case η_k is quite low, which is consistent with the effect of perturbations on real networks in Section 4. Further, we observe that the variation in η_k is much smoother for $k > 1$, which is not the case of the networks in Section 4. This non-monotone variation of η_k with the sampling probability is explained to some extent through Lemma 2; by analyzing η_k in the Erdős-Rényi model, we show rigorously that this is an inherent aspect of most graphs.

Lemma 2. *There exist constants c and pairs (p_1, p_2), where $0 < p_1 < p_2 < 1$ such that for $G \in \mathcal{G}(n, c/n)$, $\eta_1(G, G_{p_1}) > \eta_1(G, G_{p_2})$* **whp**.

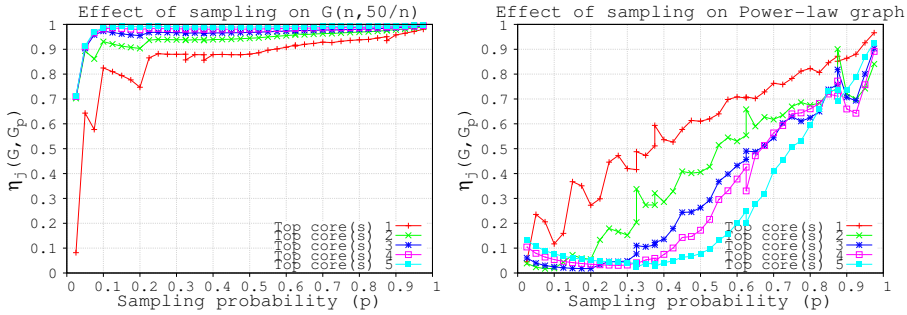

Fig. 5. Non-monotonicity of top-cores in random graphs: (a) Erdős-Rényi model $\mathcal{G}(n, c/n)$, with $n = 10000$ and $c = 50$; (b) Chung-Lu model with $n = 10000$ and average degree 5.

This lemma relies on the result of [19]. Suppose $G \in \mathcal{G}(n, \lambda/n)$. Let $Po(\mu)$ denote a Poisson random variable with mean μ. For a positive integer j, let $\psi_j(\mu) := P\left(Po(\mu) \geq j\right)$ and let $\lambda_j := \min_{\mu>0} \mu/\psi_{j-1}(\mu)$. Let, for $\lambda > \lambda_j$, $\mu_j(\lambda) > 0$ denote the largest solution to $\mu/\psi_{j-1}(\mu) = \lambda$. Pittel et al. [24] show that if $\lambda < \lambda_k$ and $k \geq 3$, then k-core $C_k(G)$ is empty **whp**, while if $\lambda > \lambda_k$, $|C_k(G)| = \psi_k(\mu_k(\lambda))n$, **whp**.

Proof. (of Lemma 2) First we note that for $G \in \mathcal{G}(n, c/n)$, the random graph sampled with probability p, G_p itself is a $\mathcal{G}(n, cp/n)$ random graph. We choose $c = 50$, for which **whp** $k_{\max}(G) = 38$ and $|C_{k_{\max}}(G)| \approx 0.91n$. We set $p_1 = 0.102$, such that cp_1 is slightly less than $\lambda_4 \approx 5.15$ (in the context of the result of [24]). For this p_1, $k_{\max}(G_{p_1}) = 3$ and $|C_{k_{\max}}(G_{p_1})| \geq 0.864n$ **whp**. We choose $p_2 = 0.198$, such that cp_2 is slightly more than $\lambda_7 \approx 9.88$. This means $k_{\max}(G_{p_2}) = 7$ and $|C_{k_{\max}}(G_{p_2})| \approx 0.694n$ **whp**. Now we show that $\eta_1(G, G_{p_1}) > \eta_1(G, G_{p_2})$ **whp**.

For any set U and subsets $A, B \subseteq U$, the following inequality follows trivially: $\frac{|A|+|B|-|U|}{|U|} \leq JI(A, B) \leq \frac{|B|}{|A|}$. We set $A = C_{k_{\max}}(G)$ and $U = V(G)$. Setting $B = C_{k_{\max}}(G_{p_1})$ and using the lower bound in the above inequality, the Jaccard Index at p_1 is $\geq 0.91 + 0.864 - 1 = 0.774$. Setting $B = C_{k_{\max}}(G_{p_2})$ and using the upper bound in the above inequality, the Jaccard Index at p_2 is $\leq 0.694/0.91 \approx 0.762$. Hence, proved. □

Remark 1. The proof of Lemma 2 is a rigorous analysis of the non-monotone behavior seen in Figure 5(a) in the interval $[0.1, 0.2]$. Similar pairs can be demonstrated for other values of c which correspond to $k_{\max} = 39, 40$ and so on.

7 Conclusions

Our results show that the top cores show significant sensitivity to perturbations, and can be recovered to a reasonable extent in sampled graphs, only if the sampling rate is sufficiently high. These results suggest that a careful sensitivity analysis is necessary when using the core structure, especially because of the non-monotone effects on the similarity index of the top cores. Our formulation

of the CorePerturbation problem and its #P-hardness implies quantifying the effects of uncertainty can be a challenging problem even in very simplified noise models; developing efficient algorithms for this problem is an interesting open problem. Further, the non-monotone behavior in the similarity index of the top cores implies simple statistical tests that might try to improve the confidence by bounding the uncertainty might not work. The reduced non-monotonicity in η_k with k suggests considering the top few cores, instead of just the top core, as a way of dealing with these effects; however, as we observe, this would require considering a much larger set of nodes. The significant sensitivity to sampling also suggests the need for greater care in the use of networks inferred using small samples provided by public APIs of social media applications. We expect our approach to be useful in the analysis of the sensitivity of other network properties to noise and sampling.

Acknowledgments. We are grateful to the reviewers whose comments have helped improve the paper. This work has been partially supported by the following grants: DTRA Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0010, NSF Career CNS 0845700, NSF ICES CCF-1216000, NSF NETSE Grant CNS-1011769 and DOE DE-SC0003957. Also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.
Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

1. D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling. *J. ACM*, 56(4):21:1–21:28, July 2009.
2. A. Adiga and A. Vullikanti. How robust is the core of a network? `http://ndssl.vbi.vt.edu/supplementary-info/vskumar/kcore.pdf`.
3. José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. K-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *NHM*, 3(2):371–393, 2008.
4. Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.
5. S. Borgatti, K. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28:124–136, 2006.
6. Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.

7. F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6:125–145, 2002.
8. A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.
9. E. Costenbader and T. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25:283–307, 2003.
10. SN Dorogovtsev, AV Goltsev, and JFF Mendes. k-core architecture and k-core percolation on complex networks. *Physica D: Nonlinear Phenomena*, 224(1):7–19, 2006.
11. N. Dronen. PyHRG. `https://github.com/ndronen/PyHRG`, 2013.
12. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, volume 29, pages 251–262, 1999.
13. D Fernholz and V Ramachandran. Cores and connectivity in sparse random graphs. Technical report, UTCS TR04-13, 2004.
14. Abraham D Flaxman and Alan M Frieze. The diameter of randomly perturbed digraphs and some applications. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 345–356. Springer, 2004.
15. A.D. Flaxman. Expansion and lack thereof in randomly perturbed graphs. *Internet Mathematics*, 4(2-3):131–147, 2007.
16. Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the Twitterers-predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, pages 3–3. USENIX Association, 2010.
17. Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1, 2011.
18. Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J Honey, Van J Wedeen, and Olaf Sporns. Mapping the structural core of human cerebral cortex. *PLoS biology*, 6(7):e159, 2008.
19. S. Janson and M.J. Luczak. A simple solution to the k-core problem. *Random Structures & Algorithms*, 30(1-2):50–62, 2007.
20. Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.
21. J. Leskovec. Stanford network analysis project. `http://snap.stanford.edu/index.html`, 2011.
22. F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.
23. M. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
24. Boris Pittel, Joel Spencer, and Nicholas C. Wormald. Sudden emergence of a giantk-core in a random graph. *J. Comb. Theory, Ser. B*, 67(1):111–151, 1996.
25. D. Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends in TCS*, 5(2):73205, 2010.
26. David J Schwab, Robijn F Bruinsma, Jack L Feldman, and Alex J Levine. Rhythmogenic neuronal networks, emergent leaders, and k-cores. *Physical Review E*, 82(5):051911, 2010.
27. D. Spielman. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, pages 76–84, 2009.