Soc. Netw. Anal. Min. (2013) 3:829–842
DOI 10.1007/s13278-013-0096-x

ORIGINAL ARTICLE

Robustness of social and web graphs to node removal

Paolo Boldi · Marco Rosa · Sebastiano Vigna

Received: 30 May 2012 / Revised: 11 November 2012 / Accepted: 10 January 2013 / Published online: 29 January 2013

Abstract Given a social network, which of its nodes have a stronger impact in determining its structure? More precisely, which node-removal order has the greatest impact on the network structure? We approach this well-known problem for the first time in a setting that combines both web graphs and social networks. Our experiments are performed on datasets that are of orders of magnitude larger than those appearing in the previous literature: this is possible, thanks to some recently developed algorithms and software tools that approximate accurately the number of reachable pairs and the distribution of distances in large graphs. Our experiments highlight deep differences in the structure of social networks and web graphs, show significant limitations of previous experimental results; at the same time, they reveal *clustering by label propagation* as a new and very effective way of locating nodes that are important from a structural viewpoint.

Keywords Social networks · Graph mining · Clustering · Graph centrality

1 Introduction

In the last years, there has been an ever-increasing research activity in the study of real-world complex networks (the world-wide web, the Internet autonomous-systems graph, co-authorship graphs, phone-call graphs, email graphs and biological networks, to cite but a few). These networks, typically generated directly or indirectly by human activity and interaction, appear in a large variety of contexts and often exhibit a surprisingly similar structure. One of the most important notions that researchers have been trying to capture in these graphs is "node centrality": ideally, every node (often representing an individual) has some degree of influence or importance within the social domain under consideration, and one expects such importance to be reflected in the structure of the social network; centrality is a quantitative measure that aims at revealing the importance of a node.

Among the types of centrality that have been considered in the literature (Borgatti (2005) for a good survey), many have to do with the shortest paths between nodes; for example, the *betweenness centrality* of a node v is the sum, over all pairs of nodes x and y, of the fraction of the shortest paths from x to y passing through v. The role played by the shortest paths is justified by one of the most well-known features of complex networks: the so-called small-world phenomenon.

A small-world network (Cohen and Havlin 2010) is a graph where the average distance between nodes is logarithmic in the size of the network, whereas the clustering coefficient is larger (that is, neighbourhoods tend to be denser) than in a random Erdős-Rényi graph with the same size and average distance.[1] Here, and in the following, by "distance" we mean the length of the shortest path between two nodes. The fact that the social networks (either electronically mediated or not) exhibit the small-world

Paolo Boldi and Sebastiano Vigna have been partially supported by a Yahoo! faculty grant, by MIUR (Italian Ministry of University and Research) and by the EU-FET grant NADINE (GA 288956).

P. Boldi · M. Rosa · S. Vigna (✉)
Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
e-mail: vigna@dsi.unimi.it

[1] The reader might find this definition a bit vague, and some variants are often spotted in the literature: this is a general problem, also highlighted recently by Li et al. (2005).

property is known at least since Milgram's famous experiment (Travers and Milgram 1969),[2] and is arguably the most popular of all features of complex networks.

Based on the above observation that the small-world property is by far the most crucial of all the features that the social networks exhibit, it is quite natural to consider centrality measures that are based on node distance, like betweenness. On the other hand, albeit interesting and profound, such measures are often computationally very expensive to be actually computed on real-world graphs. For example, the best-known algorithm to compute betweenness centrality (Brandes 2001) takes time $O(nm)$ and requires space for $O(n + m)$ integers (where n is the number of nodes and m is the number of arcs): both bounds are infeasible for large networks, where one can have $n \approx 10^9$ and $m \approx 10^{11}$. For this reason, in most cases, other strictly local measures of centrality are usually preferred (e.g., degree centrality).

One of the ideas that have emerged in the literature is that the node centrality can be evaluated based on how much the removal of a node "disrupts" the graph structure (Albert et al. 2000). This idea provides also a notion of robustness of the network: if removing few nodes has no noticeable impact, then the network structure is clearly robust in a very strong sense. On the other hand, a node-removal strategy that quickly affects the distribution of distances probably reflects an importance order of the nodes.

Previous literature has used mainly the diameter or some analogous measure to establish whether the network structure changed. Recently, though, there have been some successful attempts to produce reliable estimates of the *neighbourhood function* of very large graphs (Palmer et al. 2002; Boldi et al. 2011a), an immediate application of these approximate algorithms is the computation of the number of *reachable pairs* of the graph and its *distance distribution*.[3] The techniques used to compute distance distributions can be actually adapted to compute quickly and accurately a number of known measures (e.g., closeness centrality; Bavelas 1950) and of new ones. An example of one such new measure is *harmonic centrality* (Boldi and Vigna 2012b), defined on a node x by

$$h(x) = \sum_{y \neq x} \frac{1}{d(y,x)},$$

that is, the sum of the reciprocals of all distances to the node; this summation is extended to all $y \neq x$, with the proviso that the infinite distances give a null contribution (i.e., $1/\infty = 0$). Harmonic centrality is actually proportional to the reciprocal of the harmonic mean of the distances $d(y, x)$, and its definition was inspired by the notion of harmonic diameter described by Marchiori and Latora (2000).

Harmonic centrality takes into account, in a natural way at the same time, the average distance of x from the other nodes and the number of nodes that can actually reach x. Although an in-depth study of harmonic centrality is not an immediate goal of this paper, we shall use it as a further structural (global) information about a network.

The considerations above lead us to focus on the following kind of experiment. We consider a certain ordering of the nodes of a graph (that is supposed to represent their "importance" or "centrality"). We remove nodes (and of course their incident arcs) following this order, until a certain fraction ϑ of the arcs have been deleted.[4] At the end, we compare the resulting graph with the original one, to see how much they differ. The chosen ordering is considered to be a reliable measure of centrality if the measured difference increases rapidly with ϑ: it is sufficient to delete a small fraction of important nodes to change the structure of the graph. The comparison between the two graphs (the original one and the one obtained after node removal) is performed based on the number of reachable pairs and on the distance distribution among them.

In this work, we applied the described approach to various complex networks, considering different orderings, and obtained the following results:

- In all complex networks we considered, the removal of a limited fraction of *randomly chosen* nodes does not change the distance distribution significantly, confirming previous results.
- In web graphs, URL depth (i.e., distance from the site root) is a good measure of importance; removing homepages largely disrupts the distance distribution.
- We tested strategies based on PageRank, clustering, harmonic and betweenness centrality (see Sect. 4.1 for more information about this), and showed that they (in particular, the last three) disrupt quickly the structure of a web graph.
- Maybe surprisingly, none of the above strategies seem to have an impact when applied to social networks

[2] It should be remarked that the Milgram's experiment tried to prove two properties at the same time. First, the average distance between individuals is much smaller than expected; second, the individuals are able to exploit such a feature to route messages along short paths, albeit they only possess local information about the network they live in. This second property is, in a sense, not only more interesting than the former, but also more difficult to describe and study, because it has to do with some information that the nodes possess about the environment they inhabit.

[3] A reachable pair is a pair of nodes $\langle x, y \rangle$ such that there is a directed path from x to y; the distance distribution of a graph is a discrete distribution that gives, for every t, the fraction of reachable pairs of nodes that are at distance t.

[4] Observe that we delete nodes but count the percentage of arcs (rather than nodes) that have been removed: this choice is justified by the fact that otherwise node orderings that put large-degree nodes first would certainly be considered (unfairly) more disruptive.

other than web graphs. This is yet another example of a profound structural difference between web graphs and social networks,[5] on the same line as those discussed in Boldi et al. (2011a) and Chierichetti et al. (2009). This observation, in particular, suggests that the social networks tend to be more robust and cohesive than the web graphs, at least as far as distances are concerned; moreover, they show that "scale-free" models, which are currently proposed for both type of networks, do not to capture this important difference.

2 Related works

The idea of grasping information about the structure of a network by repeatedly removing nodes out of it is not new: Albert et al. (2000) study experimentally the variation of the diameter on two different models of *undirected* random graphs when nodes are removed either randomly or in "connectedness order" and report different behaviors. They also perform tests on some small real dataset, and we will compare their results with ours in Sect. 6.

More recently, node-centrality measures that look at how some graph invariants change when some vertices or edges are deleted (sometimes called "vitality" (Brandes and Erlebach 2005b) or "induced" measures) have been studied; for example; in Borgatti (2006) (identifying nodes that maximally disconnect the network) or in Borgatti et al. (2006) (related to the uncertainty of data).

Donato et al. (2008) study how the size of the giant component changes when nodes of high indegree or outdegree are removed from the graph. While this is an interesting measure, it does not provide information about what happens outside the component.

Finally, Fogaras (2003) considers how the *harmonic diameter*[6] (the harmonic mean of the distances) changes as nodes are deleted from a small (<1 million node) snapshot of the `.ie` domain, reporting a large increase (100 %) when as little as 1,000 nodes with high PageRank are removed. The harmonic diameter is estimated by a small number of visits, however, which gives no statistical guarantee on the accuracy of the results.

Our study is very different. First of all, we use graphs that are of two orders of magnitude larger than those considered in (Albert et al. 2000) or (Fogaras 2003); moreover, we study the impact of node removal on the whole spectrum of distances. Second, we apply the removal procedures to large social networks (previous literature used only web or Internet graphs), and the striking difference in behavior shows that "scale-free" models fail to capture essential differences between these kind of networks and web graphs. Third, we document in a reproducible way all our experiments, which have provable statistical accuracy.

3 Computing the distance distribution

Given a directed graph G, its *neighbourhood function* $N_G(t)$ gives for each $t \in \mathbf{N}$ the number of pairs of nodes $\langle x, y \rangle$ such that the y is reachable from x in no more than t steps. From the neighbourhood function, several interesting features of a graph can be estimated; in this paper, we are especially interested in the *distance distribution* of the graph G, represented by the cumulative distribution function $H_G(t)$: this distribution gives the fraction of reachable pairs at distance at most t, that is, $H_G(t) = N_G(t)/\max_t N_G(t)$. The corresponding probability-density function will be denoted by $h_G(-)$. Clearly, the distance distribution contains a big deal of *global* information about the graph: graph density, average shortest-path length, diameter and effective diameter, and so on can all be obtained from the distance distribution.

Palmer et al. (2002) proposed an algorithm to *approximate* the neighbourhood function, named ANF; the authors distribute an associated tool, `snap`, which can approximate the neighbourhood function of medium-sized graphs. Before ANF, essentially no data-mining tool was able to approximate the neighbourhood function of large graphs reliably. A remarkable exception is Cohen's work (Cohen 1997), which provides strong theoretical guarantees but experimentally turns out to be not as scalable as the ANF approach; it is worth noting, though, that one of the proposed applications (*On-line estimation of weights of growing sets*) of (Cohen 1997) is structurally identical to ANF.

Recently, HyperANF (Boldi et al. 2011a) emerged as an evolution of ANF. HyperANF can compute for the first time in a few hours the neighbourhood function of graphs with billions of nodes with a small error and good confidence using a standard workstation. HyperANF keeps track of the number of nodes reachable from each node using *HyperLogLog counters* (Flajolet et al. 2007), a kind of

[5] We remark that several proposals have been made to find features that highlight such structural differences in a computationwise-feasible way (e.g., assortative mixing (Newman and Park 2003)), but all instances we are aware of have been questioned by the subsequent literature, so no clear-cut results are known so far. An exception is the idea of considering the *spid* [shortest-path index of dispersion (Boldi et al. 2011a)], which is experimentally larger than the one for web graphs and smaller than the one for social networks. For instance, the spid of the entire Facebook graph is 0.09 (Backstrom et al. 2012).

[6] Actually, the notion had been introduced before by Marchiori and Latora (2000) and named *connectivity length*, but we find the name "harmonic diameter" much more insightful.

sketch that makes it possible to compute the number of distinct elements of a stream in very little space; such counters can be thought as dictionaries that can answer just questions about size: the answer is probabilistic and depends on a random seed that is chosen independently for each run. Each counter is made of a number of small *registers*, and the precision of the answer depends on the number of registers.

The free availability of HyperANF opens new and interesting ways to study large graphs, of which this paper is an example. HyperANF was also successfully employed in the first world-scale social-network graph-distance computations, using the entire Facebook network of active users ($\approx$721 million users, $\approx$69 billion friendship links), determining an average distance of 4.74 (Backstrom et al. 2012).

4 Removal strategies and their analysis

In the previous section, we discussed how we can effectively approximate the distance distribution of a given graph G; we shall use such a distribution as the graph structural property of interest.

Consider some total order $\prec$; on the nodes of G; we think of $\prec$ as a removal strategy in the following sense: when we want to remove ϑm arcs, we start removing the $\prec$-largest node (and its incident arcs), go on removing the second-$\prec$-largest node, etc., and stop as soon as $\geq \vartheta m$ arcs have been removed. The resulting graph will be denoted by $G(\prec, \vartheta)$. Of course, $G(\prec, 0) = G$ whereas $G(\prec, 1)$ is the empty graph. We are interested in measuring how different $G(\prec, \vartheta)$ is from G: looking at how this measure of difference changes when ϑ varies, we can judge the ability of $\prec$ to identify nodes that will disrupt the network. The measures of difference we shall consider are all based on global properties. In particular, we will consider the following differences: (a) the change in the fraction of reachable pairs; (b) the *divergence*[7] between the distribution H_G and the distribution $H_{G(\prec,\vartheta)}$.

4.1 Some removal strategies

We considered several different strategies for removing nodes from a graph. Some of them embody actually significant knowledge about the structure of the graph, whereas others are very simple (or even independent of the graph) and will be used as baseline. Some of them have been proposed in the previous literature, and will be useful to compare our results.

As a first observation, some strategies requires a *symmetric* (a.k.a. *undirected*) graph: in this case, we symmetrise the graph by adding the missing arcs[8] The second obvious observation is that some strategies might depend on available metadata (e.g., URLs for web graphs) and do not make sense for all graphs.

Random No strategy: we pick random nodes and remove them from the graph. It is important to test against this "nonstrategy" as we can show that the phenomena we observe are due to the peculiar choice of nodes involved, and not to some generic property of the graph.

Largest-degree first We remove nodes in decreasing (out- or in-)degree order. This strategy is an obvious baseline, as *degree centrality* is the first shot at centrality in a network.

Near-root In web graphs, we can assume that nodes that are roots of websites and their (quasi-)immediate successors (e.g., pages linked by the root) are most important in establishing the distance distribution, as people tend to link higher levels of websites. This strategy removes essentially root nodes first, then the nodes that are children of a root on, and so on.

PageRank PageRank (Page et al. 1998) is a well-known algorithm that assigns ranks to nodes using a Markov chain obtained from the graph. It has been designed as an improvement over degree centrality, because nodes with high degree which, however, are connected to nodes of low rank will have a rather low rank (the definition is indeed recursive). There is a vast body of literature on the subject see Boldi et al. (2009) and Langville and Meyer (2004), and the references therein.

Label propagation Label propagation (Raghavan et al. 2007) is a powerful technique for clustering symmetric graphs.[9] Each node has a label (initially, the node number itself) and through a number of rounds, each node changes its label by taking the label of the majority of its neighbours. At the end, node labels are used as cluster identifiers. Our removal strategy picks first, for each cluster in decreasing size order, the node with the highest number of neighbours in other clusters: intuitively, it is a representative of a set of tightly connected nodes (the cluster) which, however, has a very significant connection with the outside world (the other clusters) and, thus, we expect that its removal should seriously disrupt the distance distribution.

[7] We purposely use the word "divergence" between distributions, instead of "distance", to avoid confusion with the notion of distance in a graph.

[8] It is mostly a matter of taste whether to use directed symmetric graphs or simple undirected graphs. In our case, since we have to cope with both directed and undirected graph, we prefer to speak of directed graphs that are symmetric, that is, for every arc $x \to y$, there is a symmetric arc $y \to x$.

[9] Label propagation has been independently proposed under the name of *peer pressure clustering* by Gilbert et al. (2007).

Once we have removed all such nodes, we proceed again, cluster by cluster, using the same criterion (thus, picking the second node of each cluster that has more connection toward other clusters), and so on.

Betweenness centrality The *betweenness centrality* (Anthonisse 1971; Freeman 1977) of a node v is the sum, over all pairs of nodes x and y, of the fraction of shortest paths from x to y passing through v. Betweenness centrality is difficult to compute, as it requires [using the algorithm described by Brandes (2001)] n breadth-first visits. We used a highly parallel implementation of Brandes' algorithm which enabled us to compute betweenness centrality on our two smallest examples (a social network and a web graph).[10]

Harmonic centrality Finally, we can employ harmonic centrality (Boldi and Vigna 2012b) as a removal strategy, removing the nodes with the largest centrality first. We recall that the harmonic centrality of a node x is the sum of the reciprocals of the distances between every other node y and x.

4.2 Measures of divergence

Once we changed the structure of a graph by deleting some of its nodes (and arcs), there are several ways to measure whether the structure of the graph has significantly changed. The first, basic raw datum we consider is the change in the number of pairs of nodes that are still reachable.

Then, we observe the change in the shape of the distance distribution (comparing the distribution in the modified graph with one of the original graph). In the top row of Figs. 1 and 2, we show how the distribution changes for four different graphs (described in Sect. 5) using the label-propagation strategy; the figure presents the probability mass function of the distance distribution for different values of ϑ. The reader can see that, for web graphs, the distribution changes shape and its mode moves to the right (witnessing the fact that shortest paths tend to get longer as we keep removing arcs), and at some point (when $\vartheta \geq 0.2$), the change in shape is radical and the distribution has virtually no relation with the original one. The phenomenon is much less evident on social networks.

To compare quantitatively two distributions, we considered various measure of divergence; in particular, we considered the following possibilities (here, P denotes the original distance distribution, and Q the distribution after node removal):

Relative average-distance change This is somehow the simplest and most natural measure: how much has the average distance between reachable pairs changed? We use the measure

$$\delta(P, Q) = \frac{\mu_Q}{\mu_P} - 1$$

where μ denotes the average; in other words, we measure how much the average value changed. This measure is non-symmetric, but it is of course easy to obtain $\delta(P, Q)$ from $\delta(Q, P)$.

Relative harmonic-diameter change This measure is analogous to the relative average-distance change, but the average on distances is *harmonic* and *computed on all pairs*, that is:

$$\frac{n(n-1)}{\sum_{x \neq y} \frac{1}{d(x,y)}} = n(n-1) \Big/ \sum_{t>0} \frac{1}{t} (N_G(t) - N_G(t-1)),$$

where n is the number of nodes of the graph. This measure, proposed by Marchiori and Latora (2000) and used by Fogaras (2003), includes reachability information, as unreachable pairs contribute zero to the sum. It is easily computable from the neighbourhood function, as shown above.

ℓ norms. A further alternative is given by viewing distance distributions as functions $\mathbf{N} \to [0 \ldots 1]$ and measure their distance using some ℓ-norm, most notably ℓ_1 or ℓ_2. Such distances are of course symmetric.

We tested these divergences with various graphs and removal strategies, to understand how the choice of distribution divergence influences the interpretation of the results obtained. In the bottom rows of Figs. 1 and 2, we plot the outcomes, but the results are consistent in all the cases we tested. Note that the figures for divergences in web graph had to be split into two, because the range of the change in harmonic diameter is much wider than any other measure.

We strive for a measure that *increases monotonically* as more and more nodes are removed from the network. This is a somehow basic requirement—a measure that fluctuates when we try to increasingly disconnect a network is not measuring what we are interested in. Moreover, we expect the range of the measure to be related to the strength of the change.

Most of the times, all measures agree (apart for obvious scale factors), at least for $\vartheta < 0.2$. Note that in some cases (e.g., the Hollywood graph), the range of variation is within the precision of our approximate computation: in this case, observing fluctuations is normal. Nonetheless, there are a number of obvious pathological behaviors suggesting that a number of measures do not satisfy our criteria.

Change of the fraction of reachable pairs In the case of the social network `orkut`-2007, then the number of reachable pairs is essentially unchanged even when $\vartheta = 0.3$; nonetheless, the structure of the network has changed, as the increase in average distance shows.

[10] There are sampled variants of Brandes's algorithm (Brandes and Erlebach 2005a), but the Hoeffding bound providing precision guarantees requires $\Theta(n^4 \log n/\varepsilon^2)$ visits to obtain absolute precision ε.

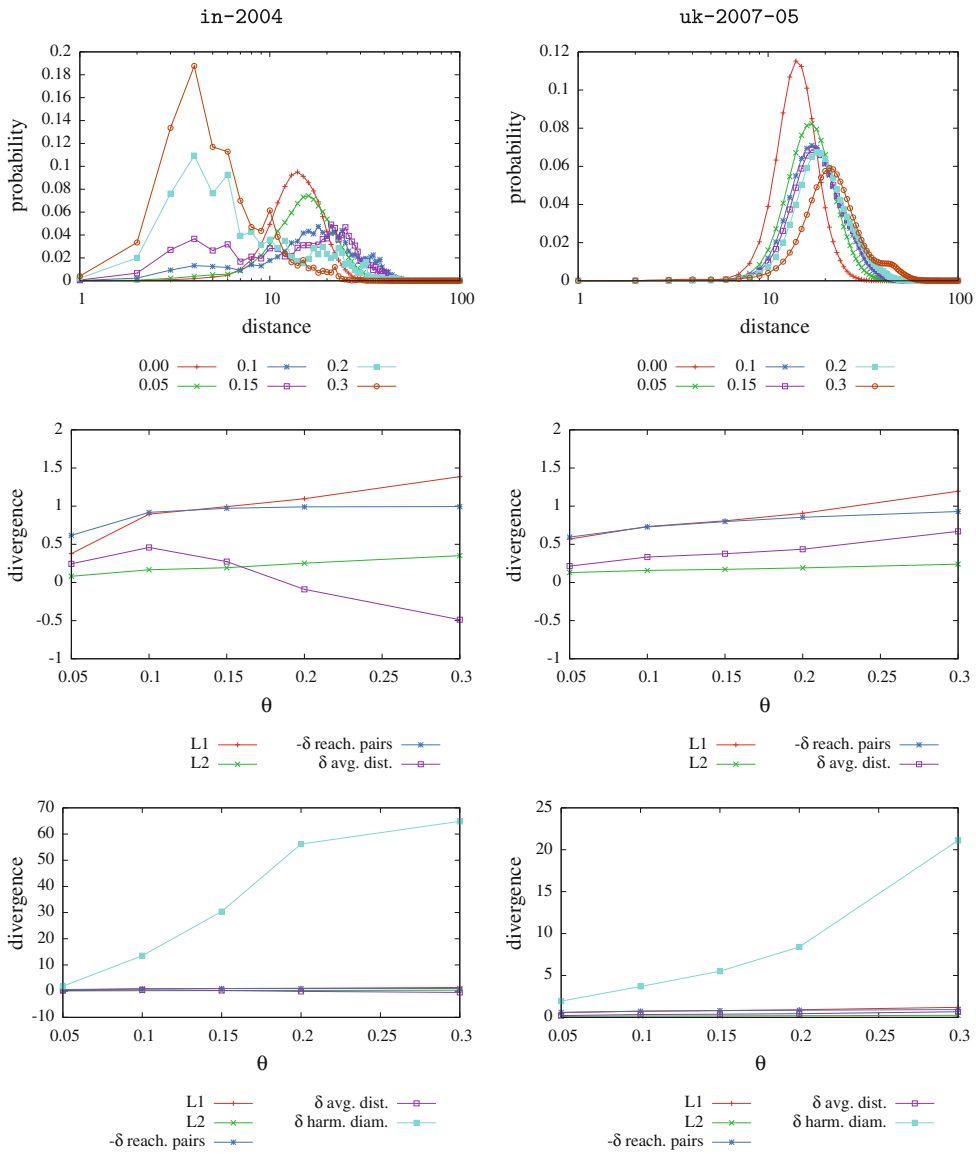

Fig. 1 Testing various divergence measures on two web graphs under the label-propagation strategy: in the *left column*, a small 2004 snapshot of the `.in` domain, and in the *right column* a larger 2007 snapshot of the `.uk` domain. At the *top*, we show how the distance distribution changes for different values of ϑ; then, we show the behavior of all divergence measures, except for the change in harmonic diameter; finally, at the *bottom*, we show all divergences measures

Relative average-distance change In the case of the web graph in-2004, when ϑ gets large we observe that the the δ average distance decreases; this apparently strange phenomenon has a rather simple explanation: the shortest paths get longer at the beginning due to the removal of arcs, but when the percentage of removed arcs becomes very large, the graph becomes more disconnected, and existing shortest paths start getting shorter. This fact (see also Boldi and Vigna 2012a) suggest that this measure is not useful when networks get significantly disconnected.

ℓ norms The ℓ_2 norm has a very small variation on the web graph `uk`-2007-05, in spite of a significant variation of the average distance and of a large variation of the number of reachable pairs; finally, the ℓ_1 norm has essentially the same range of variation for `uk`-2007-05 and `orkut`-2007, even if the changes in the first network are much more significant.

All in all, we conclude that the *change in harmonic diameter* is the most reliable measure of connectness of a network, confirming the intuition of Marchiori and Latora (2000). While analyzing single aspect (fraction of reachable pairs, etc.) is obviously useful to understand changes at a finer level of detail, the harmonic diameter provides a compact representation of the changes in the structure of the network that keeps track both of disconnection and of changes in the distance distribution.

5 Experiments

For our experiments, we considered a number of networks with various sizes and characteristics; most of them are either web graphs or (directed or undirected) social graphs of some kind (note that for web graphs, we can rely on the

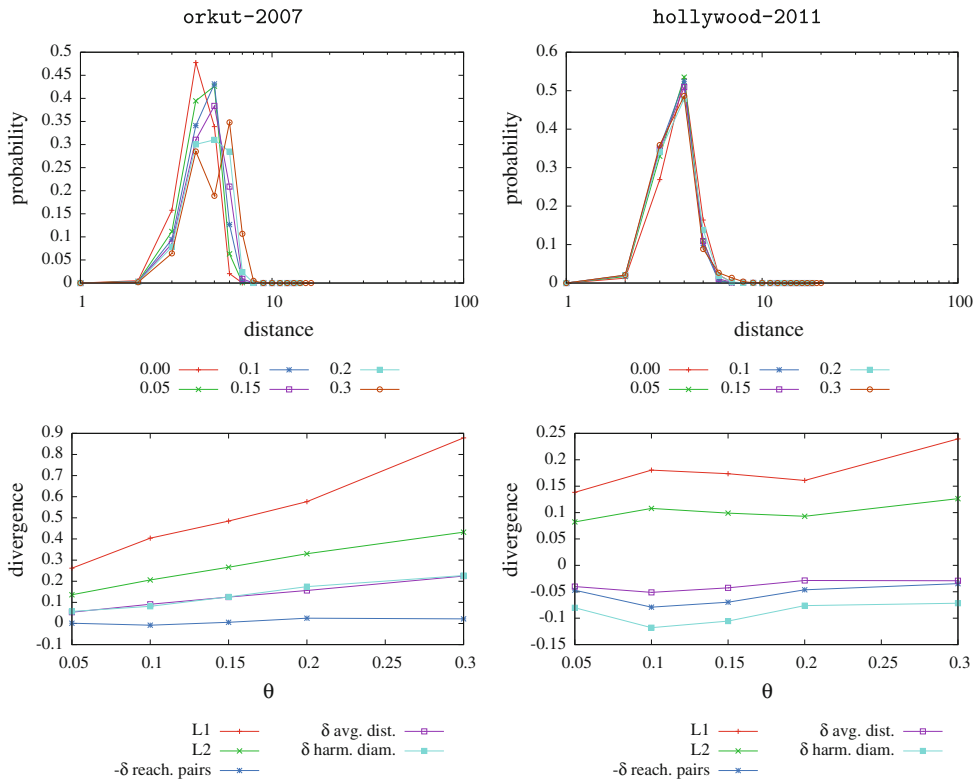

Fig. 2 Testing various divergence measures on two social networks under the label-propagation strategy: in the *left column*, a 2007 snapshot of the Orkut social network, and in the *right column* a 2011 snapshot of the Hollywood co-starring network. At the *top*, we show how the distance distribution changes for different values of ϑ; then, we show the behavior of all divergence measures

URLs as external source of information). In this paper, we are going to present the results about the following datasets:[11]

- *Hollywood* (1,985,306 nodes, 114,492,816 undirected edges): One of the most popular *undirected* social graphs, the graph of movie actors: vertices are actors, and two actors are joined by an edge whenever they appeared in a movie together.
- *LiveJournal* (5,363,260 nodes, 79,023,142 directed arcs): LiveJournal is a virtual community social site started in 1999: nodes are users and there is an arc from x to y if x registered y among his friends (it is not necessary to ask y permission, so the graph is *directed*). We considered the same 2008 snapshot of *LiveJournal* used by Chierichetti et al. (2009) for their experiments.
- *Orkut* (3,072,626 nodes, 117,185,083 undirected edges): Orkut was a social networking and discussion site operated by Google. This snapshot is a part of the IMC 2007 Data Sets (Mislove et al. 2007).
- For comparison, we considered two web graphs of different size: a small 2004 snapshot of the `.in` domain ($\approx$1.3 million nodes), and a snapshot taken in May 2007 of the `.uk` domain ($\approx$100-million nodes).

We remark that all our graphs are available at the LAW website.[12] HyperANF is available as free software at the WebGraph website,[13] and the class RemoveHubs that has been used to perform the experiments we describe is part of the LAW software.

We applied our removal strategies with different impact levels ϑ (i.e., percentage of removed arcs), namely 0.05, 0.1, 0.15, 0.2 and 0.3. For each level, we ran HyperANF at least ten times using 1,024 registers per counter for all networks (except .uk, for which we used 512-register counters due to its large size); this setting guarantees that the HyperANF estimates the number of nodes at any given distance with a relative standard deviation that never exceeds 1.45 %.

Tables 1, 2 and 3 show how the harmonic diameter, average distance and percentage of reachable pairs change for the different datasets and strategies considered. In the tables, we are reporting the jackknife (Efron and Gong 1983) estimate of derived values (such as average distances or harmonic diameter) and the associated estimation of the

[11] In Boldi et al. (2011b), we also presented the outcomes of similar experiments performed on other networks (Amazon, Enron and `.it`) that agree with the ones shown here. Our tables and graphs slightly differ from those previously published in (Boldi et al. 2011b), because we had time to generate more runs, and thus, increase the precision of our results.

[12] http://law.di.unimi.it/. In particular, the graphs we used are the datasets named hollywood-2011, ljournal-2008, orkut-2007, in-2004 and uk-2007-05. Note that isolated nodes have been removed from hollywood-2011.

[13] http://webgraph.di.unimi.it/.

standard error. Observe that the latter (that we may consider as a measurement error) is essentially negligible, and will, therefore, be ignored when data are being compared.

The relative change in the harmonic diameter is shown in Table 4; the same values for two of the datasets are plotted in Fig. 3, along with the relative average-distance change and the percentage of reachable pairs.

6 Discussion

Let us start our discussion by looking at Table 2 showing the average distance: as we anticipated in Sect. 4.2, we observe almost always that this quantity increases with ϑ (because deleting arcs tends to make the shortest paths longer); sometimes, though, (especially when ϑ is large and "good" strategies like LP are used) there is a drop, due to the fact that some pairs become disconnected, hence, paradoxically reducing the average distance. This fact is better understood if one compares Table 2 with Table 3, which shows the percentage of reachable pairs.

Table 1, reporting the harmonic diameter, is our main source of information: here, the two effects (disconnection and change of distribution) are combined, and we observe a constant increase in the harmonic diameters for all strategies and all ϑ; note how dramatic this increase appears to be in some cases. The changes are better read in Table 4 that reports the δ between the harmonic diameter in the modified graph and the original one.

Table 1 For each graph and fractions of removed arcs, we show the harmonic diameter along with the estimation of the standard error in the measurement obtained by the jackknife. PR stands for PageRank, HC for harmonic centrality and LP for label propagation

Graph	Strategy	0.05	0.1	0.15	0.2	0.3
Hollywood 4.05 (±0.04)	Random	4.08 (±0.03)	4.11 (±0.04)	4.12 (±0.04)	4.05 (±0.03)	4.23 (±0.04)
	Degree	4.08 (±0.03)	4.12 (±0.04)	4.20 (±0.04)	4.24 (±0.04)	4.40 (±0.04)
	PR	4.14 (±0.05)	4.17 (±0.04)	4.22 (±0.04)	4.25 (±0.04)	4.47 (±0.03)
	HC	4.08 (±0.04)	4.20 (±0.03)	4.23 (±0.04)	4.32 (±0.05)	4.60 (±0.05)
	LP	3.73 (±0.05)	3.58 (±0.03)	3.63 (±0.03)	3.75 (±0.03)	3.76 (±0.03)
	Betweeness	4.10 (±0.04)	4.20 (±0.03)	4.30 (±0.03)	4.42 (±0.03)	4.58 (±0.05)
LiveJourna 7.36 (±0.07)	Random	7.54 (±0.06)	7.76 (±0.07)	7.92 (±0.07)	8.14 (±0.07)	8.29 (±0.05)
	Indegree	7.53 (±0.10)	7.74 (±0.07)	7.90 (±0.06)	7.97 (±0.09)	8.67 (±0.06)
	Outdegree	7.50 (±0.04)	7.76 (±0.11)	7.92 (±0.06)	8.27 (±0.06)	8.57 (±0.08)
	PR	7.61 (±0.05)	7.95 (±0.07)	8.22 (±0.06)	8.61 (±0.05)	9.32 (±0.13)
	HC	7.69 (±0.08)	8.00 (±0.05)	8.33 (±0.09)	8.72 (±0.11)	9.63 (±0.11)
	LP	7.49 (±0.05)	7.39 (±0.06)	7.27 (±0.08)	7.23 (±0.06)	7.68 (±0.06)
Orkut 4.06 (±0.01)	Random	4.07 (±0.04)	4.09 (±0.03)	4.10 (±0.04)	4.13 (±0.02)	4.24 (±0.04)
	Degree	4.21 (±0.04)	4.33 (±0.04)	4.36 (±0.02)	4.39 (±0.05)	4.61 (±0.04)
	PR	4.24 (±0.05)	4.36 (±0.06)	4.51 (±0.04)	4.64 (±0.05)	4.80 (±0.04)
	HC	4.28 (±0.05)	4.36 (±0.04)	4.40 (±0.04)	4.51 (±0.03)	4.68 (±0.03)
	LP	4.29 (±0.03)	4.39 (±0.03)	4.57 (±0.04)	4.76 (±0.03)	4.98 (±0.06)
.in 32.26 (±0.24)	Random	36.53 (±0.33)	39.59 (±0.38)	45.57 (±0.43)	54.62 (±0.35)	75.22 (±0.76)
	Indegree	38.44 (±0.51)	47.03 (±0.40)	57.74 (±0.30)	68.21 (±0.54)	87.68 (±0.78)
	Outdegree	36.85 (±0.29)	37.82 (±0.16)	37.99 (±0.42)	38.38 (±0.37)	50.91 (±0.38)
	Near-Root	181.18 (±1.77)	239.18 (±2.02)	284.04 (±2.26)	352.80 (±2.78)	1,021.47 (±3.91)
	PR	44.61 (±0.22)	58.16 (±0.45)	82.81 (±0.90)	130.36 (±1.47)	330.23 (±2.61)
	HC	73.40 (±0.69)	143.56 (±0.70)	357.12 (±2.26)	941.41 (±6.21)	1,475.91 (±5.02)
	LP	93.03 (±0.62)	469.13 (±4.68)	1,012.98 (±5.88)	1,844.96 (±3.85)	2,124.97 (±5.25)
	Betweeness	263.90 (±2.18)	3,125.36 (±6.75)	6,222.17 (±13.35)	8,966.32 (±18.09)	15,073.35 (±18.78)
.uk 22.78 (±0.24)	Random	24.43 (±0.20)	26.37 (±0.37)	28.24 (±0.32)	31.01 (±0.47)	36.13 (±0.46)
	Indegree	25.40 (±0.57)	27.89 (±0.30)	30.76 (±0.40)	33.27 (±0.37)	45.65 (±0.49)
	Outdegree	23.28 (±0.33)	23.91 (±0.36)	24.30 (±0.37)	25.07 (±0.39)	28.13 (±0.41)
	Near-Root	55.22 (±0.59)	55.73 (±0.95)	59.68 (±1.21)	64.67 (±0.89)	80.44 (±1.06)
	PR	30.29 (±0.35)	36.39 (±0.43)	44.22 (±0.46)	50.92 (±0.55)	73.04 (±0.94)
	HC	30.27 (±0.40)	41.00 (±0.54)	57.89 (±0.46)	90.89 (±1.10)	240.90 (±1.84)
	LP	66.71 (±0.97)	106.75 (±1.27)	148.26 (±2.40)	214.12 (±2.00)	503.89 (±7.97)

Table 2 For each graph and fractions of removed arcs, we show the average distance along with the standard error in the measurement obtained by the jackknife. PR, HC and LP have the same meaning as in Table 1

Graph	Strategy	0.05	0.1	0.15	0.2	0.3
Hollywood 3.92 (±0.00)	Random	3.92 (±0.01)	3.94 (±0.00)	3.95 (±0.01)	3.96 (±0.01)	3.97 (±0.01)
	Degree	3.97 (±0.01)	4.02 (±0.01)	4.06 (±0.00)	4.12 (±0.00)	4.23 (±0.00)
	PR	3.99 (±0.01)	4.03 (±0.01)	4.10 (±0.00)	4.15 (±0.01)	4.26 (±0.00)
	HC	3.99 (±0.00)	4.04 (±0.01)	4.09 (±0.01)	4.16 (±0.00)	4.31 (±0.01)
	LP	3.76 (±0.00)	3.72 (±0.00)	3.75 (±0.01)	3.81 (±0.01)	3.80 (±0.01)
	Betweenness	4.02 (±0.00)	4.11 (±0.01)	4.18 (±0.00)	4.24 (±0.01)	4.44 (±0.01)
LiveJournal 5.99 (±0.01)	Random	6.02 (±0.01)	6.01 (±0.01)	6.04 (±0.01)	6.06 (±0.01)	6.12 (±0.01)
	Indegree	6.05 (±0.01)	6.15 (±0.01)	6.23 (±0.01)	6.32 (±0.01)	6.55 (±0.01)
	Outdegree	6.10 (±0.01)	6.17 (±0.01)	6.27 (±0.01)	6.36 (±0.01)	6.60 (±0.01)
	PR	6.10 (±0.01)	6.23 (±0.01)	6.36 (±0.01)	6.50 (±0.01)	6.87 (±0.01)
	HC	6.19 (±0.01)	6.35 (±0.01)	6.49 (±0.01)	6.66 (±0.01)	7.05 (±0.01)
	LP	5.86 (±0.00)	5.82 (±0.00)	5.82 (±0.00)	5.85 (±0.01)	6.03 (±0.01)
Orkut 4.21 (±0.00)	Random	4.22 (±0.00)	4.24 (±0.00)	4.25 (±0.00)	4.27 (±0.00)	4.31 (±0.01)
	Degree	4.38 (±0.01)	4.43 (±0.00)	4.47 (±0.00)	4.53 (±0.01)	4.67 (±0.01)
	PR	4.40 (±0.00)	4.51 (±0.01)	4.57 (±0.00)	4.62 (±0.00)	4.75 (±0.01)
	HC	4.39 (±0.00)	4.47 (±0.01)	4.53 (±0.01)	4.59 (±0.01)	4.74 (±0.01)
	LP	4.44 (±0.00)	4.60 (±0.00)	4.74 (±0.01)	4.87 (±0.01)	5.16 (±0.01)
.in 15.34 (±0.04)	Random	15.20 (±0.03)	15.57 (±0.04)	15.64 (±0.03)	15.46 (±0.04)	15.96 (±0.03)
	Indegree	15.78 (±0.02)	16.11 (±0.03)	16.92 (±0.04)	16.99 (±0.04)	18.98 (±0.27)
	Outdegree	17.69 (±0.06)	18.48 (±0.15)	18.62 (±0.34)	18.32 (±0.22)	19.33 (±0.20)
	Near-Root	22.99 (±0.07)	22.82 (±0.04)	23.01 (±0.05)	23.44 (±0.05)	15.94 (±0.05)
	PR	16.17 (±0.04)	16.50 (±0.06)	17.93 (±0.04)	20.98 (±0.26)	32.45 (±0.76)
	HC	21.95 (±0.05)	26.22 (±0.10)	27.44 (±0.08)	38.55 (±0.32)	13.66 (±0.04)
	LP	19.10 (±0.36)	22.39 (±0.06)	19.53 (±0.07)	13.95 (±0.13)	7.86 (±0.02)
	Betweeness	28.04 (±0.09)	26.40 (±0.08)	29.94 (±0.25)	42.02 (±0.38)	71.37 (±0.64)
.uk 15.42 (±0.04)	Random	15.66 (±0.03)	15.80 (±0.04)	16.05 (±0.03)	16.21 (±0.06)	16.62 (±0.04)
	Indegree	16.15 (±0.07)	16.41 (±0.03)	16.76 (±0.12)	17.11 (±0.05)	18.06 (±0.05)
	Outdegree	15.59 (±0.08)	15.54 (±0.04)	15.94 (±0.17)	15.95 (±0.05)	16.98 (±0.07)
	Near-Root	18.93 (±0.04)	19.03 (±0.09)	19.16 (±0.05)	20.87 (±1.08)	18.74 (±0.15)
	PR	16.49 (±0.05)	17.05 (±0.03)	17.66 (±0.11)	18.14 (±0.06)	19.40 (±0.05)
	HC	18.38 (±0.04)	20.86 (±0.05)	23.63 (±0.09)	27.11 (±0.09)	36.09 (±0.48)
	LP	18.73 (±0.03)	20.55 (±0.09)	21.22 (±0.04)	22.13 (±0.03)	25.74 (±0.07)

A first, clear remark that all these data consistently show is that the social networks suffer spectacularly less disconnection than the web graphs when their nodes are removed using our strategies (see Fig. 4). Our two most efficient removal strategies, label propagation and betweenness, can disconnect almost all pairs of a web graph by removing less than 20 % of the arcs, whereas they does not affect much the percentage of reachable pairs on social networks. This entirely different behavior shows that the web graphs have a path structure that passes through fundamental hubs, something that does not seem to take place in social networks.

Moreover, the harmonic diameter of web graphs we consider can almost be doubled by removing only 5 % of the arcs, and increases by as much as 60 times upon the removal of 30 % of the arcs. In most social networks, there is just an increase of a few percents (in any case, always less than 6 %).[14] This is also very clear looking at Fig. 3.

Note that random removal can separate a good number of reachable pairs (Table 3), but the increase in average distance is very marginal (Table 2). This shows again that considering both measures is important in evaluating removal strategies.

Of course, we cannot state that there is no strategy able to disrupt social networks as much as a web graph (simply

[14] We remark that in some cases, the measure is negative or does not decrease monotonically. This is sometimes an artifact of the probabilistic technique used to estimate our measures—small relative errors are unavoidable.

Table 3 For each graph and fractions of removed arcs, we show the percentage of reachable pairs along with the standard error in the measurement obtained by the jackknife. PR, HC and LP have the same meaning as in Table 1

Graph	Strategy	0.05	0.1	0.15	0.2	0.3
Hollywood 92.92 (±1.01)	Random	92.44 (±0.67)	92.01 (±1.05)	92.13 (±1.00)	94.06 (±0.85)	90.11 (±0.96)
	Degree	93.63 (±0.81)	93.84 (±0.90)	93.18 (±0.97)	93.54 (±0.99)	92.79 (±0.76)
	PR	92.57 (±1.11)	93.09 (±1.00)	93.62 (±0.95)	94.01 (±1.07)	92.08 (±0.75)
	HC	94.07 (±0.93)	92.71 (±0.65)	93.38 (±0.87)	93.01 (±1.16)	90.50 (±1.03)
	LP	97.28 (±1.32)	100.26 (±0.90)	99.39 (±0.99)	97.22 (±0.80)	96.12 (±0.75)
	Betweenness	94.51 (±0.93)	94.08 (±0.80)	93.39 (±0.71)	92.33 (±0.61)	93.17 (±1.18)
LiveJournal 78.62 (±0.78)	Random	77.00 (±0.68)	74.77 (±0.67)	73.62 (±0.73)	71.90 (±0.69)	71.25 (±0.51)
	Indegree	77.71 (±1.19)	76.76 (±0.74)	76.35 (±0.62)	76.74 (±0.92)	73.25 (±0.53)
	Outdegree	78.57 (±0.50)	76.86 (±1.18)	76.58 (±0.65)	74.41 (±0.62)	74.53 (±0.70)
	PR	77.55 (±0.57)	75.79 (±0.71)	74.88 (±0.65)	73.12 (±0.43)	71.48 (±1.06)
	HC	77.85 (±0.85)	76.70 (±0.58)	75.49 (±0.89)	74.06 (±1.01)	70.94 (±0.84)
	LP	75.74 (±0.57)	76.21 (±0.70)	77.56 (±0.92)	78.38 (±0.76)	75.88 (±0.71)
Orkut 100.00 (±0.29)	Random	100.13 (±1.05)	100.12 (±0.71)	100.12 (±0.93)	99.74 (±0.59)	97.96 (±1.12)
	Degree	100.84 (±1.05)	99.35 (±1.04)	99.53 (±0.59)	100.16 (±1.34)	98.31 (±0.99)
	PR	100.52 (±1.27)	100.15 (±1.42)	98.09 (±0.91)	96.44 (±1.12)	95.78 (±0.83)
	HC	99.28 (±1.15)	99.10 (±1.09)	99.72 (±1.01)	98.61 (±0.85)	97.84 (±0.82)
	LP	99.87 (±0.79)	100.82 (±0.87)	99.43 (±0.96)	97.48 (±0.67)	97.80 (±1.31)
.in 43.30 (±0.39)	Random	37.88 (±0.39)	35.43 (±0.37)	30.77 (±0.31)	25.55 (±0.18)	18.90 (±0.22)
	Indegree	37.26 (±0.52)	30.96 (±0.31)	26.49 (±0.14)	22.60 (±0.21)	18.62 (±0.21)
	Outdegree	38.53 (±0.33)	37.73 (±0.18)	37.57 (±0.42)	37.10 (±0.32)	28.98 (±0.26)
	Near-Root	10.46 (±0.13)	7.75 (±0.08)	6.42 (±0.06)	5.01 (±0.05)	0.91 (±0.01)
	PR	32.60 (±0.19)	25.62 (±0.23)	19.48 (±0.23)	13.46 (±0.17)	6.34 (±0.07)
	HC	26.46 (±0.31)	15.53 (±0.09)	6.07 (±0.05)	2.00 (±0.02)	0.64 (±0.00)
	LP	16.55 (±0.14)	3.50 (±0.05)	1.23 (±0.01)	0.39 (±0.00)	0.24 (±0.00)
	Betweenness	8.79 (±0.10)	0.34 (±0.00)	0.10 (±0.00)	0.05 (±0.00)	0.03 (±0.00)
.uk 63.55 (±0.74)	Random	60.09 (±0.54)	56.08 (±0.87)	53.08 (±0.68)	48.71 (±0.85)	42.78 (±0.60)
	Indegree	59.54 (±1.52)	55.10 (±0.65)	50.67 (±0.65)	48.00 (±0.63)	36.67 (±0.42)
	Outdegree	62.52 (±0.97)	60.92 (±0.94)	60.35 (±0.95)	59.17 (±0.97)	54.41 (±0.82)
	Near-Root	31.28 (±0.38)	31.09 (±0.59)	28.74 (±0.63)	26.76 (±0.39)	21.16 (±0.32)
	PR	50.83 (±0.65)	43.61 (±0.50)	36.85 (±0.43)	33.01 (±0.39)	24.56 (±0.36)
	HC	57.28 (±0.86)	47.93 (±0.70)	38.19 (±0.41)	27.67 (±0.40)	13.15 (±0.13)
	LP	25.79 (±0.39)	17.29 (±0.24)	12.85 (±0.23)	9.25 (±0.09)	4.45 (±0.08)

because, this strategy may be different from the ones that we considered), but the fact that all strategies work very similarly in both cases (e.g., label propagation is by far the most disruptive strategy) suggests that the phenomenon is intrinsic.

There is a candidate easy explanation: the shortest paths in web graphs pass frequently through home pages, which are linked more than other pages. But this explanation does not take into account the fact that clustering by label propagation and betweenness centrality are significantly more effective than the Near-root removal strategy. Rather, it appears that there are fundamental hubs (not necessarily home pages) which act as shortcuts and through which a large number of the shortest paths pass. Label propagation and betweenness centrality are able to identify such hubs, and their removal results in an almost disconnected graph and in a very significant increase in average distance.

These hubs are not necessarily of high in- or out-degree: quite the opposite, rather, is true. The behavior of web graphs under the largest-degree strategy is illuminating: we obtain a small reduction in reachable pairs, an almost unnoticeable change of the average distance and a very marginal one for the harmonic diameter: all these facts together suggest that nodes of high degree are not actually so relevant for the global structure of the network.

Social networks are much more resistant to node removal. There is no strict clustering or definite hubs, which can be used to eliminate or elongate the shortest paths. This is perhaps explainable since networks emerging from social interaction are much less engineered (there is

Table 4 For each graph and fractions of removed arcs, we show the relative harmonic diameter change (the δ measure between the harmonic diameter in the modified graph and the original one). PR, HC and LP have the same meaning as in Table 1

Graph	Strategy	0.05	0.1	0.15	0.2	0.3
Hollywood	Random	0.005	0.014	0.015	−0.002	0.044
	Degree	0.006	0.017	0.035	0.046	0.084
	PR	0.022	0.029	0.040	0.049	0.102
	HC	0.006	0.035	0.043	0.065	0.135
	LP	−0.080	−0.118	−0.105	−0.076	−0.072
	Betweenness	0.010	0.035	0.061	0.091	0.129
LiveJournal	Random	0.025	0.055	0.077	0.106	0.127
	Indegree	0.023	0.052	0.073	0.083	0.178
	Outdegree	0.020	0.055	0.076	0.124	0.165
	PR	0.034	0.081	0.117	0.171	0.267
	HC	0.045	0.088	0.132	0.186	0.309
	LP	0.018	0.005	−0.012	−0.017	0.044
Orkut	Random	0.003	0.008	0.010	0.019	0.046
	Degree	0.038	0.066	0.074	0.082	0.137
	PR	0.046	0.075	0.111	0.143	0.182
	HC	0.055	0.076	0.084	0.111	0.154
	LP	0.057	0.081	0.126	0.174	0.227
.in	Random	0.132	0.227	0.413	0.693	1.331
	Indegree	0.192	0.458	0.790	1.114	1.718
	Outdegree	0.142	0.172	0.177	0.190	0.578
	Near-Root	4.615	6.413	7.804	9.935	30.660
	PR	0.383	0.803	1.567	3.040	9.235
	HC	1.275	3.449	10.069	28.178	44.745
	LP	1.883	13.540	30.396	56.183	64.862
	Betweenness	7.179	95.868	191.851	276.903	466.185
.uk	Random	0.072	0.158	0.240	0.361	0.586
	Indegree	0.115	0.224	0.350	0.460	1.004
	Outdegree	0.022	0.050	0.066	0.100	0.235
	Near-Root	1.424	1.446	1.620	1.838	2.531
	PR	0.330	0.597	0.941	1.235	2.206
	HC	0.328	0.800	1.541	2.989	9.573
	LP	1.928	3.685	5.507	8.398	21.116

no notion of "site" or "page hierarchy", for example) than web graphs.

Comparing the strategies with one another, it seems clear that the more powerful are label propagation and betweenness (and, for web graphs, also Near-root), with harmonic centrality as a close contender. For small values of ϑ, Near-root, when applicable, is very effective, but it is soon overtaken by harmonic centrality, label propagation and betweenness, the latter being by far the most powerful altogether. This shows that while the removal of root, pages has an initial powerful effect, removing pages at higher levels has no longer a significant impact.

How are the rankings provided by the best techniques correlated? Surprisingly, very little. We computed Kendall's τ (Kendall 1945) on the rankings given by Near-root order, label propagation, harmonic centrality and betweenness centrality on the `.in` snapshot and on the Hollywood graph. The absolute value of τ is always below 0.12 (almost complete uncorrelation), the only exception being a value of 0.39, when comparing harmonic and betweenness centrality on the Hollywood graph.

It is interesting to compare our results with those in the previous literature. With respect to (Albert et al. 2000), we tested much larger networks. We can confirm that the random removal is less effective than the rank-based removal, but clearly the variation in diameter measured in (Albert et al. 2000) has been made on a *symmetrized* version of the web graph. Symmetrization destroys much of the structure of the network, and it is difficult to justify (you cannot navigate links backwards). We have evaluated our experiment using the variation in the diameter instead of the variation in average distance (not shown here), but

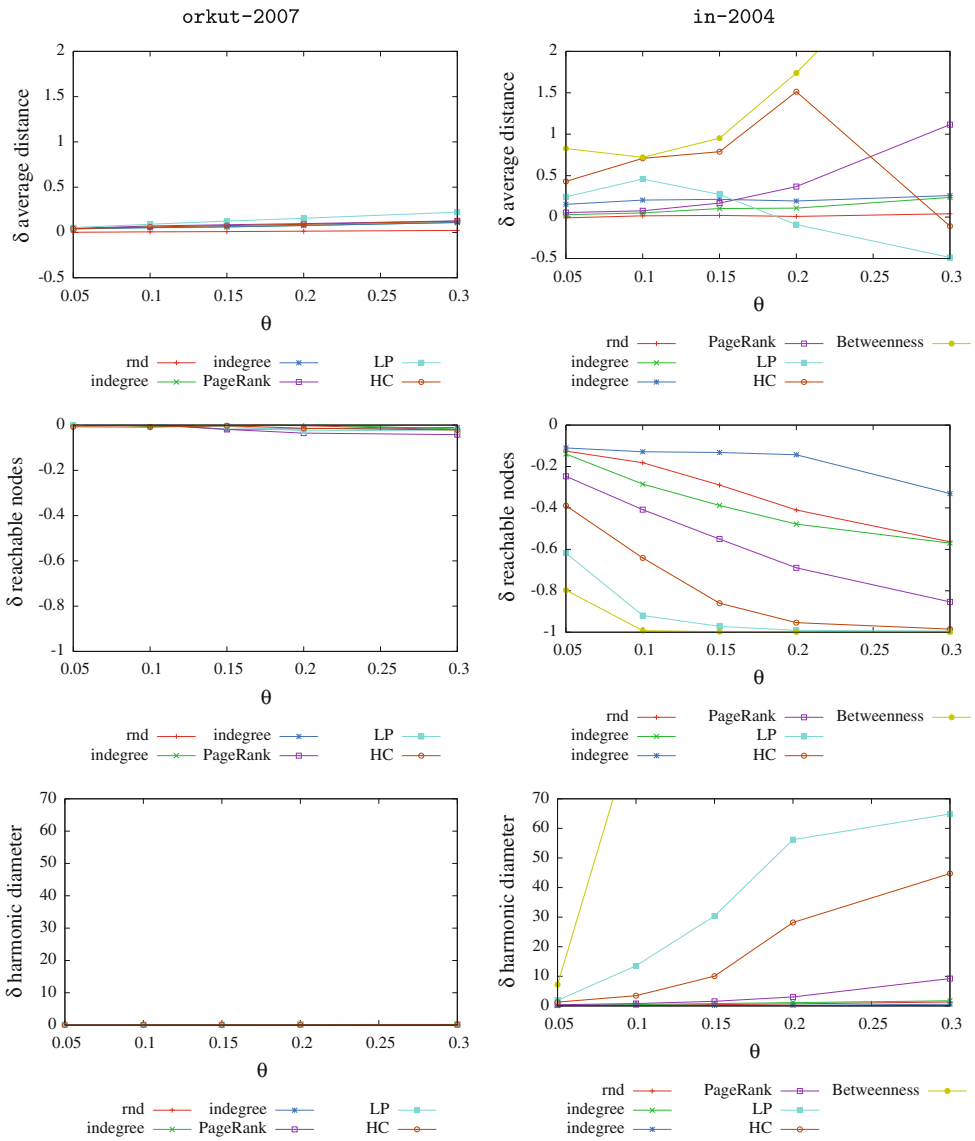

Fig. 3 Typical behavior of social networks (Orkut, *left*) and web graphs (`.in`, *right*) when a ϑ fraction of arcs is removed using various strategies. We purposely show the two plots using the same range for the y axis, to highlight how none of the proposed strategies completely disrupts the structure of social networks; conversely, the effect of some strategies on web graphs (especially, the label-propagation removal strategy) is very visible

the results are definitely inconclusive. The behavior is wildly different even between graphs of the same type, and shows no clear trend. This was expected, as the diameter is defined by a maximization property, so it is very unstable.

Evaluating the variation in harmonic diameter allows us to compare our data with those of Fogaras (2003): as we already remarked, the harmonic diameter is very interesting, because it combines reachability and distance. The data confirm what we already stated: web graphs react to removal of 30 % of their arcs through label propagation by increasing dramatically their harmonic diameter—something that does not happen with social networks.

Our criterion for node elimination is a threshold on the number of *arcs* removed, rather than nodes, so a strictly numerical comparison of our results with that of Fogaras (2003) is not possible. However, for `.uk` PageRank at $\vartheta = 0.01$ removes 648 nodes, which produced in the `.ie` graph a relative increment of 100 %, whereas we find 14 %. This is to be expected, due to the very small size of the dataset used in (Fogaras 2003): experience shows that connectedness phenomena in web graphs are very different in the "below ten million nodes" region (e.g., see the different behavior of our `.in` dataset). Nonetheless, the growth trend is visible in both cases. However, the experiments in Fogaras (2003) fail to detect both the disruptive behavior at $\vartheta = 0.3$ and the striking difference between largest-degree and PageRank strategy.

7 Conclusions and future work

We have explored experimentally the alterations of the distance distribution of some social networks and web graphs under different node-removal strategies. We have

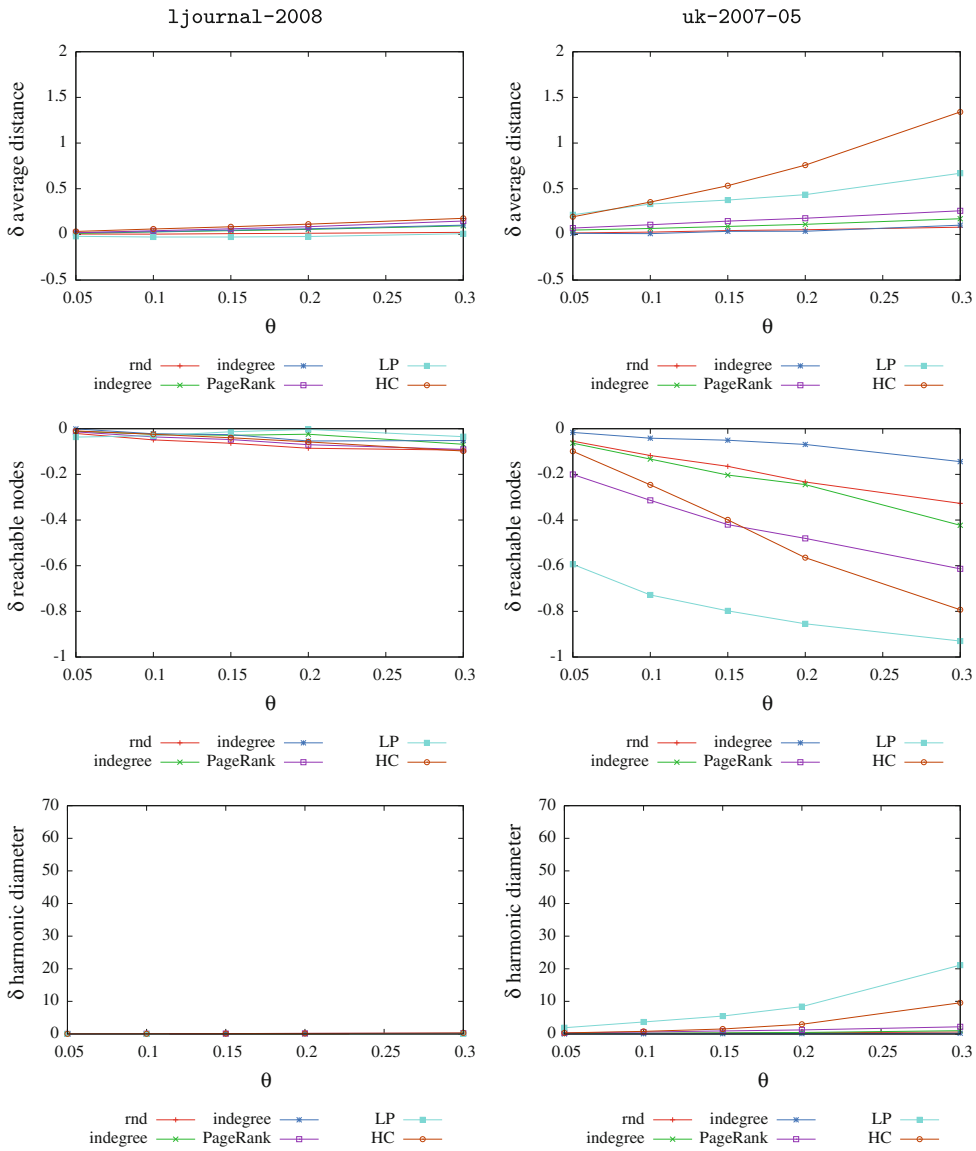

Fig. 4 Typical behavior of social networks (LJournal, *left*) and web graphs (.uk, *right*) when a ϑ fraction of arcs is removed using various strategies. The range is fixed, and the same of Fig. 4. Note the more regular behavior of the .uk snapshot with respect to the smaller .in snapshot in Fig. 4. Note also that on the .uk snapshot harmonic centrality increases more the average distance, but label propagation makes more pairs unreachable

confirmed some of the experimental results that appeared in the literature, but at the same time, shown some basic limitations of previous approaches. In particular, we have shown for the first time that there is a clear-cut structural difference between social networks and web graphs,[15] and that it is important to test node-removal strategies until a significant fraction of the arcs have been removed.

Probably the most important conclusion is that "scale-free" models, which are currently proposed for both web graphs and social networks, do not to capture this important difference: for this reason, they can only make sense as long as they are adopted as baselines.

It would be extremely interesting, though, to find analytical tools that allow one to approach the structure change (i.e., to see what impact a given removal strategy has on a given network) in a more analytical way: such tools would be necessary to design new probabilistic network models that behave like real-world social networks do according to our experiments.

It might be argued that reachable pairs and distance distributions are too coarse as a feature. Nonetheless, we believe that they are the most immediate *global* features that are approachable computationally. For instance, checking whether node removal alters the clustering coefficient would not be so interesting, because the clustering coefficient of each node depends only on the structure of its very neighbourhood. Thus, by removing first the nodes with high coefficient, it would be trivial to make the clustering coefficient of the graph decrease quickly. Such

[15] In this paper, like in all the other experimental research on the same topic, conclusions about social networks should be taken with a grain of salt, due to the heterogeneity of such networks and the lack of a large repertoire of examples.

trivial approaches cannot possibly work with reachable pairs or with distance distributions, because they are properties that depend on the graph *as a whole*.

References

Albert R, Jeong H, Barabási A-L (2000) Error and attack tolerance of complex networks. Nature 406:378–382

Anthonisse JM (1971) The rush in a graph. Technical report, University of Amsterdam Mathematical Centre, Amsterdam

Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S (2012) Four degrees of separation. In: ACM Web Science 2012: Conference Proceedings, pp 45–54 (Best paper award)

Bavelas A (1950) Communication patterns in task-oriented groups. J Acoust Soc Am 57:271–282

Boldi P, Vigna S (2012a) Four degrees of separation, really. Arxiv preprint arxiv:1205.5509

Boldi P, Vigna S (2012b). Harmonic centrality (in preparation)

Boldi P, Santini M, Vigna S (2009) Page Rank: functional dependencies. ACM Trans Inf Sys 27(4):1–23

Boldi P, Rosa M, Vigna S (2011a) HyperANF: approximating the neighbourhood function of very large graphs on a budget. In: Srinivasan S, Ramamritham S, Kumar A, Ravindra MP, Bertino E, Kumar R (eds) Proceedings of the 20th international conference on World Wide Web. ACM, pp 625–634

Boldi P, Rosa M, Vigna S (2011b) Robustness of social networks: Comparative results based on distance distributions. In: Social Informatics, Third International Conference, SocInfo 2011. Lecture Notes in Computer Science, vol 6894. Springer, Berlin, pp 8–21

Borgatti SP (2005) Centrality and network flow. Soc Netw 27(1): 55–71

Borgatti SP (2006) Identifying sets of key players in a social network. Comput Math Organ Theory 12:21–34

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28(2):124–136

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177

Brandes U, Erlebach T (eds) (2005a) Network analysis: methodological foundations. Lecture Notes in Computer Science, vol 3418. Springer, Berlin

Brandes U, Erlebach T (2005b) Network analysis: methodological foundations (Lecture Notes in Computer Science). Number 3418 in Lecture Notes in Computer Science. Springer, Berlin

Chierichetti F, Kumar R, Lattanzi S, Mitzenmacher M, Panconesi A, Raghavan P (2009) On compressing social networks. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 219–228

Cohen E (1997) Size-estimation framework with applications to transitive closure and reachability. J Comput Syst Sci 55:441–453

Cohen R, Havlin S (2010) Complex networks: structure, robustness and function. Cambridge University Press, Cambridge

Donato D, Leonardi S, Millozzi S, Tsaparas P (2008) Mining the inner structure of the web graph. J Phys A: Math Theor 41(22): 224017

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37(1):36–48

Flajolet P, Fusy É, Gandouet O, Meunier F (2007) HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In: Proceedings of the 13th conference on analysis of algorithm (AofA 07), Juan-les-Pins, pp 127–146

Fogaras D (2003) Where to start browsing the web? In: Innovative Internet Community Systems, Third International Workshop, IICS 2003. Lecture Notes in Computer Science, vol 2877. Springer, Leipzig, pp 65–79

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41

Gilbert JR, Reinhardt S, Shah V (2007) High-performance graph algorithms from parallel sparse matrices. In: Kagstrom B, Elmroth E, Dongarra J, Wasniewski J (eds) Applied parallel computing. State of the Art in Scientific Computing (8th PARA'06). Lecture Notes in Computer Science, vol 4699. Springer, New York, pp 260–269

Kendall MG (1945) The treatment of ties in ranking problems. Biometrika 33(3):239–251

Langville AN, Meyer CD (2004) Deeper inside Page Rank. Internet Math 1(3):355–400

Li L, Alderson DL, Doyle J, Willinger W (2005) Towards a theory of scale-free graphs: definition, properties, and implications. Internet Math 2(4):431–523

Marchiori M, Latora V (2000) Harmony in the small-world. Physica A: Stat Mech Appl 285(3-4):539–546

Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07), San Diego

Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68(3):036122

Page L, Brin S, Motwani R, Winograd T (1998) The Page Rank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project, Stanford University, Stanford

Palmer CR, Gibbons PB, Faloutsos C (2002) Anf: a fast and scalable tool for data mining in massive graphs. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, pp 81–90

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106. doi:10.1103/PhysRevE.76.036106

Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32(4):425–443