Efficient Measurement of Complex Networks Using Link Queries

Fabien Tarissan, Matthieu Latapy, Christophe Prieur

Efficient Measurement of Complex Networks Using Link Queries

Fabien Tarissan
ISC
CNRS and École Polytechnique
tarissan@lix.polytechnique.fr

Matthieu Latapy
LIP6
CNRS and Université Pierre et Marie Curie
Matthieu.Latapy@lip6.fr

Christophe Prieur
LIAFA
Université Paris Diderot
prieur@liafa.jussieu.fr

***Abstract*—Complex networks are at the core of an intense research activity. However, in most cases, intricate and costly measurement procedures are needed to explore their structure. In some cases, these measurements rely on link queries: given two nodes, it is possible to test the existence of a link between them. These tests may be costly, and thus minimizing their number while maximizing the number of discovered links is a key issue. This is a challenging task, though, as initially no information is known on the network. This paper studies this problem: we observe that properties classically observed on real-world complex networks give hints for their efficient measurement; we derive simple principles and several measurement strategies based on this, and experimentally evaluate their efficiency on real-world cases. In order to do so, we introduce methods to evaluate the efficiency of strategies. We also explore the bias that different measurement strategies may induce.**

I. Preliminaries

Complex networks, modeled as large graphs, are everywhere in science, society, and everyday life: in transportation (airlines, roads); communication (internet, web, file or email exchanges); social life (collaborations, friendship, economical exchanges); life sciences (interactions between proteins or genes, dependencies between species); language analysis (synonymy, co-occurrences of words); etc. As a consequence, much effort is devoted to the analysis and modeling of such networks, leading to key insight on various key topics like epidemy or information spreading, algorithm and protocol design, resilience to failures and attacks, etc.

However, it must be clear that most real-world complex networks are not directly available: collecting information on their structure generally relies on intricate and expensive measurement procedures. Conducting such a measurement often is a challenge in itself, and is an important part of the work needed to study a complex network.

In general, complex network measurements consist in a combination of a few simple measurement primitives. In several cases, this primitive consists in testing the existence of a link, which we call a *link query*: given two nodes u and v, a measurement operation makes it possible to decide whether there is a link between them or not. This simple test may be expensive (regarding the needed resources or time, or the load it induces on the network, for instance) and so conducting measurements with as few calls to the measurement primitive as possible is a key issue.

For instance, in online social networks like Facebook or Flickr[1], privacy concerns and reduction of server load often lead to limitations in the queries that one is allowed to perform to explore networks between users. Link queries are however allowed in most cases. Likewise, measurements of real-world social networks often rely on interviews, in which link queries play a central role [1]. In biological networks like protein interactions or gene regulatory networks, link queries also play a key role [2], [3].

In all these contexts, and others, link queries are very expensive: they have a significant load on server running online social network software and their number is generally bounded; they have a significant cost for interviewers and participants in sociological studies ; or they require costly biological experiments, depending on the case.

In this paper, we formalise this problem as follows: given a graph $G = (V, E)$, we want to define *strategies* (ordered lists of link queries) which lead to the discovery of as many links of the network as possible. In other words, we want to minimize the number of link queries while maximizing the number of observed links, i.e. the number of positive answers to these tests[2].

In order to do so, we will rely on simple intuitions derived from statistical properties observed on most real-world complex networks, which we discuss in Section II. We then propose several measurement strategies in Section III based on these principles. We also need a way to compare and evaluate measurement strategies, see Section IV. We finally use this to experimentally evaluate proposed strategies in Section V.

Before entering in the core of this paper, we give the needed formalism and notations, and discuss related work.

A. Formalism and notations

In all the paper, we will consider an undirected[3] graph $G = (V, E)$, with $n = |V|$ nodes and $m = |E|$ links. We suppose that all the nodes are known, and focus on link discovery only. In other words, we know V but know nothing

[1] http://www.facebook.com/ and http://www.flickr.com/

[2] Notice that, whereas we suppose that link queries are very expensive, the computational cost of each strategy is not our concern here; we consider it as negligible compared to measurement costs, which fits most real-world cases.

[3] This means that we make no difference between (u, v) and (v, u), for any u and v.

about E (although we will make some statistical assumptions in accordance with classical empirical observations in the field, see Section II).

We will denote by $N(v)$ the set of neighbors of $v \in V$: $N(v) = \{u \in V,\ (u,v) \in E\}$ and by $d(v)$ its degree: $d(v) = |N(v)|$.

A measurement consists in a series of link queries, *i.e.* tests of the existence of link (u,v) for two nodes u and v in V. At a given stage in such a measurement, one has already discovered a set of links, which we will denote by $E' \subseteq E$. The set of extremities of links in E' will be denoted by $V' \subset V$. Notice that, although we know V, in general $V' \neq V$. We will also denote by n' the number of nodes in V' and m' the number of discovered links so far: $n' = |V'|$ and $m' = |E'|$. We also define $N'(v) = N(v) \cap V'$ and $d'(v) = |N'(v)|$ for all $v \in V'$. Notice that both V', E', n', m', N' and d' vary during a measurement; however, the context will make it clear which value we consider.

B. Related work

This work belongs to the fields of complex network metrology, which mostly focused on the specific case of the internet topology until now, see for instance [4]–[10]. This area of research aims mainly at evaluating the relevance of collected complex network samples and properties observed on them, and correcting these observations. Viewing the measurement as the combination of many instance of a simple primitive (link queries, here) which we want to optimize is new, and is an important contribution of this paper.

Another related problem is the one of *link prediction*: given a network in which new links may appear, one wants to predict which new links will appear in the future based on currently existing ones [11], [12]. In this context, authors use properties of the known network to infer probable future link, which is similar to what we do below in the measurement context.

II. Underlying principles

Our goal is to design measurement strategies based on link queries (test of the existence of a link between two given nodes) which will minimize the number of such queries and maximize the number of discovered links (*i.e.* the number of positive answers to these tests). In order to do so, we will rely on some simple statistical properties which are observed on most real-world complex networks [13].

A. Properties of complex networks

First, we will suppose that G is sparse: its density $\delta = \frac{2.m}{n.(n-1)}$ is very small. In other words, the probability that a link exists between two randomly chosen nodes is very small, *i.e.* a random link query will fail with high probability.

The second key property is the fact that most complex networks have a very heterogeneous degree distribution (often close to a power law). Since the degree of a node is the number of links attached to it, this means that there is a high variability between the number of links of each node (many nodes have very few links, but some have more, and even many more).

Finally, another key property is the local density: although randomly chosen nodes have a very low probability to be linked, two nodes which have a neighbor in common are linked with a much higher probability. This is generally captured by the clustering coefficient or the transitivity ratio [13]–[15], defined by:

$$\mathtt{cc}(G) = \frac{\sum_v \frac{\Delta(v)}{\vee(v)}}{n}$$

$$\mathtt{tr}(G) = \frac{3.\Delta(G)}{\vee(G)}$$

where, for each $v \in V$, $\Delta(v)$ denotes the number of triangles (sets of three nodes with three links) to which v belongs; $\vee(v) = \frac{d(v).(d(v)-1)}{2}$ denotes the number of pairs of neighbors of v; $\Delta(G) = \sum_v \Delta(v)$; and $\vee(G) = \sum_v \vee(v)$.

A classical observation in complex network studies is that both these quantities are high, at least compared to the density. In other words, if one chooses a random pair of links with an extremity in common (transitivity ratio) or a random node and two of its neighbors (clustering coefficient) then the probability that the third possible link exists is high.

B. Consequences on measurements

The properties above, observed on most real-world complex networks, have a strong impact on measurements and will play a key role here.

First, the low density of complex network implies that randomly choosing two nodes and testing the presence of a link between them is very inefficient. Notice however that, when only link queries are possible, one has no choice but to begin with a series of such random measurements. However, it must be clear that exploring a large complex network with such a strategy only is not reasonable.

Instead, the existence of nodes with degree much larger than the average may be useful for efficient measurement. Suppose that we test a random pair (u,v). The probability that it is positive (*i.e.* the link (u,v) exists) is proportional to the degree of u (resp. v). Therefore, if it exists then one may guess that u (resp. v) has a high degree, and so testing all pairs (u,w) (resp. (v,w)) for any w will probably lead to the discovery of many links. Notice that u and v play a symmetric role in this reasoning. We will call this observation the *degree principle*.

Likewise, the high local density may be used for efficient measurement: when we know that two nodes u and v have a neighbor w in common then testing pair (u,v) certainly makes sense as this link exists with high probability. We call this the *triangle principle*.

We may now turn to the definition of measurement strategies based on these principles.

III. Measurement strategies

First notice that when one starts a measurement in our framework, no link is known and we have no way to distinguish between vertices. Therefore, there is no choice but to test random pairs of nodes. We call this null strategy *random*$_k$.

Strategy 1: *random*$_k$ with k an integer.

while $m' < k$ **do**
  test a random untested pair

As soon as some links are discovered, though, one may try to design more efficient strategies. The *triangle principle* indicates that, when a $\vee$ pattern is discovered one may test the missing link in the triangle. This leads to the following strategy.

Strategy 2: $\vee$-*random*$_k$ with k an integer.

while $m' < k$ **do**
  Test a random untested pair (u, v)
  if (u, v) *exists* **then**
    Test all untested pairs (v, w), for any w in $N'(u)$
    Test all untested pairs (u, w), for any w in $N'(v)$

Applying directly the *degree principle* would lead to a strategy in which we test the pairs (u, v) for all v as soon as a random test led to the discovery of a link of u. However, the *degree principle* becomes stronger if one waits until *several* links of a node are found. We therefore propose a strategy in which a series of tests (performed according to another strategy) is followed by a use of the *degree principle* on nodes for which we discovered many links.

Strategy 3: ($\vee$-)Complete Simple — *cs*$_k$ (resp. $\vee$-*cs*$_k$) with k an integer.

Apply *random*$_k$ (resp. $\vee$-*random*$_k$)
foreach $u \in V'$ *in decreasing order of* $d'(u)$ **do**
  Test all untested pairs (u, v), for any $v \in V$

This strategy may be improved by using the links it discovers for choosing the next link queries to perform. This leads to the following strategy.

Strategy 4: ($\vee$-)Complete — c_k (resp. $\vee$-c_k) with k an integer.

Apply *random*$_k$ (resp. $\vee$-*random*$_k$)
Let $X = V'$
while X *is nonempty* **do**
  Let u in X with $d'(u)$ maximal
  Remove u from X
  Test all untested pairs (u, v), for any $v \in V$
  if (u, v) *exists and is the first link of* v *discovered* **then**
    Add v to X

One may try to use an even stronger version of the *degree principle* by noticing that the probability of a link between two nodes is even larger if *both* have a high degree. Therefore, link queries between nodes for which we already discovered many links have an even higher probability of positive outcome. This leads to the following strategy.

Strategy 5: ($\vee$-)Test-Between-Found — *tbf*$_k$ (resp. $\vee$-*tbf*$_k$) with k an integer.

Apply *random*$_k$ (resp. $\vee$-*random*$_k$)
foreach $(u, v) \in V' \times V'$ *in decreasing order of* $d'(u) + d'(v)$ **do**
  Test (u, v) if it was untested

Finally, one may try to combine the strategies above in order to improve their efficiency. Indeed, some of them use complementary principles which both help in discovering more links with less link queries. One may therefore expect even better results with combinations of them. We will therefore consider the following strategy.

Strategy 6: ($\vee$-)TBF-Complete — *tbfc*$_k$ (resp. $\vee$-*tbfc*$_k$) with k an integer.

Apply *tbf*$_k$ (resp. $\vee$-*tbf*$_k$)
Apply c_0

It must be clear that many variants and improvements of the strategies above are possible. Probably, completely different strategies may also be defined. Our goal here however is to evaluate the relevance of the *degree principle* and *triangle principle* in the design of measurement strategies. We therefore focus on these relatively simple strategies, which we consider as a natural first set of strategies derived from these basic principles.

IV. Evaluation methodology

For any measurement strategy S, let us define $m'_S(q)$ as the number of links discovered with q link queries with strategy S[4]. It must be clear that our goal, for a given q, is to design a strategy S that maximises $m'_S(q)$. Conversely, one may want to discover a given number x of links and ask for the strategy S that will minimize the q such that $m'_S(q) = x$.

However, given two numbers of queries q and r it is possible that a given strategy S discovers more links with q tests than another strategy T, while T discovers more with r tests (we will observe such a situation in Section V-B). As a consequence, it makes no sense to say that S is better than T, nor the converse; this depends on the allowed number of link queries.

Going further, one may notice that if S and T discover the same number of links after a given number q of tests, but if S discovers more links than T for any number $r < q$ of test, then it seems natural to consider that S surpasses T (it discovers the same number of links, but faster).

[4]Notice that, in practice, it is in general impossible to reach a situation where we test all pairs of nodes: $q = \frac{n.(n-1)}{2}$, or conversely where we discovered all existing links: $m'_S(q) = m$.

A simple way to formalise these intuitions is to define the *efficiency* of a strategy S for a given number of queries q as the (discrete) integral of the function m'_S from 0 to q: $\mathcal{E}_q(S) = \sum_{i=1}^{q} m'_S(i)$.

Notice that the obtained value will depend on the considered graph, and on q. It seems difficult to avoid this, as the efficiency of strategies do indeed depend on the graph under concern, and on the number of allowed link queries. We will therefore always compare strategies ran on the same graph and with the same number of link queries here.

Another weakness of this definition is that it may give any positive value for the efficiency of a strategy, making it hard to evaluate how far from the worst or best solution we are. In order to avoid this we introduce the *normalised efficiency*: $\overline{\mathcal{E}}_q(S) = \frac{\mathcal{E}_q(S) - \mathcal{E}_q(\min)}{\mathcal{E}_q(max) - \mathcal{E}_q(\min)}$ where $\min$ and $\max$ stand for the worst and best strategies, *i.e.* the ones with minimal and maximal efficiencies.

Notice that strategies $\min$ and $\max$ are easy to determine: $\min$ consists in testing pairs of nodes with no links between them as long as possible, thus $\frac{n.(n-1)}{2} - m$ times, and then performing the positive tests; conversely $\max$ consists in performing first the m positive tests. As a consequence, we can compute easily $\mathcal{E}_q(\min)$ and $\mathcal{E}_q(max)$ for any q, and thus obtain the normalized efficiency of any strategy.

The notion of normalized efficiency however remains insufficient. Indeed, as we consider sparse graphs, there are only very few positive link queries, and thus one may expect to be much closer to the $\min$ strategy than to the $\max$. As a consequence, the efficiency of any strategy will be very low.

A solution to this problem consists in comparing strategies to the random one, denoted by ran, which consists in performing link queries on random untested pairs of nodes. The expected efficiency of this strategy is easy to compute, as the probability of success of a link query is exactly the density δ; we obtain: $\mathcal{E}_q(\text{ran}) = \sum_{i=1}^{q} i \cdot \delta = \frac{q.(q+1)}{2} \cdot \delta$.

Finally, we introduce the *relative efficiency*, which indicates how a given strategy S performs compared to the random one (and the minimal and maximal ones) after q link queries: $\mathcal{R}_q(S) = \frac{\overline{\mathcal{E}}_q(S)}{\overline{\mathcal{E}}_q(\text{ran})}$.

Notice that the relative efficiency does not give a value between 0 and 1 and therefore does not have the advantage of being relatively independent from the context. However, we cannot normalize it as we would loos the benefit of the comparison to the random strategy. We will therefore use both the normalized efficiency and the relative efficiency to discuss efficiency of strategies below, and keep in mind that in any case the efficiency of a strategy depends on the graph under concern and on the number of link queries allowed. Only the full $m'_S()$ function can describe the efficiency of strategy S entirely, on a given graph.

V. Experimental evaluation

In this section, we present experiments aimed at illustrating the differences between the proposed measurement strategies, and how they may be evaluated. We first present the dataset we used, which is a typical real-world case. We then examine a typical situation and discuss the observations. We deepen this by observing the impact of the initial random period of measurement; and finally we discuss the bias that measurement strategies may induce on observed properties.

A. Dataset

We use here data on an online social network which we consider as a typical example of complex networks studied in the literature. This social network comes from the *Flickr* site, which provides facilities for publishing online photos, sharing them with others, discuss them, etc. Users may also subscribe to various interest groups and have lists of other users known as their *contacts*.

Here we used a complete measurement of *Flickr* conducted in August 2006 [16]. We considered the largest of the 72 875 groups observed then [5], which contained 31 523 members.

We then defined three different networks among these 31 523 users:

- *contact*: two users a and b are linked if a is a contact of b or b is a contact of a;
- *comment*: two users a and b are linked if a posted a comment on a photo from b or b posted a comment on a photo from a;
- *symmetric-comment*: two users a and b are linked if both a posted a comment on a photo from b and b posted a comment on a photo from a.

One may also define a *symmetric-contact* graph in which two users a and b are linked if both a is a contact of b and b is a contact of a. In order to save space, we will not consider it here. Likewise, we do not detail the features of these networks; the key point here is that they are sparse, have heterogeneous degree distributions and high clustering coefficient and transitivity ratio. To this regard, they are similar to most real-world complex networks, and so the principles discussed in II apply.

B. A typical example

Let us first try all our strategies with the same parameter $k = 1\,000$ and on the *contact* graph. We represent in Figure 1 the number $m'_S(q)$ of links discovered by each strategy S as a function of the number q of link queries performed, for q between 0 and $Q = 4.10^6$. The obtained plot is representative of what is obtained on other graphs.

This plot shows clearly that measurement strategies perform very differently, and that trying to optimize them is relevant. This is a first important result in itself. Moreover, both the *degree principle* and the *triangle principle* are useful in doing so: strategies based on each of them perform significantly better than the random strategy. However, the *degree principle* seems to be much stronger: while the improvement of $\vee$-*random* remains quite low, the improvement obtained by the *complete* strategy is huge. This is probably due to the fact that, although the clustering coefficient and transitivity ratio are much larger

[5] *FlickrCentral*, http://flickr.com/groups/central/

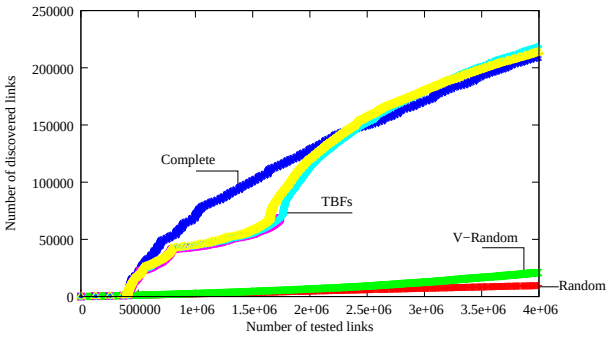

Fig. 1. Number of links (vertical axis) discovered by each strategy as a function of the number of link queries performed (horizontal axis) in a typical case (*contact* network, $k = 1\,000$). The *tbf*, *tbf-complete* and $\vee$*-tbf-complete* strategies are indicated as a unique curve (named TBFs) in the plot as the three curves overlap eachother.

than the density, they remain quite small; instead, the largest degrees in the graphs are relatively close to its number of nodes.

The best final results (the largest number of links discovered at the end of the measurement) are obtained with mixed strategies, namely *tbf-complete* and $\vee$*-tbf-complete*, which succeed in discovering between 22 and 23% of existing links by performing only 1% of possible link queries. They slightly outperform *complete*, which was expected as they are more subtle (though a stronger improvement may have been expected).

Notice that, although these strategies finally outperform *complete*, they discover links *later* than this strategy. In this sense, they may therefore be considered as less efficient, which is captured by our notion of efficiency, see Table I.

	m'	% tested	% found	$\mathcal{E}$	$\mathcal{R}$
random	9 609	1.04	1.03	0.006	0.99
$\vee$*-random*	21 030	1.04	2.25	0.010	1.64
c_{1000}	209 485	1.04	22.4	0.142	24.2
tbf_{1000}	68 874	0.46	7.36	0.048	15.6
$tbfc_{1000}$	218 448	1.04	23.4	0.131	22.3
$\vee$-$tbfc_{1000}$	214 175	1.04	22.9	0.134	22.7

TABLE I
Efficiency of each strategy after 4.10^6 links queries on the *contact* network: the number m' of discovered links; the percentage of tested pairs of nodes; the percentage of existing links found; and the efficiency coefficients $\mathcal{E}$ and $\mathcal{R}$.

C. Impact of the initial phase

In order to test the impact of the initial phase on the efficiency of the strategies, we conducted a similar experiment in which we increased the parameter k to 1 500. We present the results in Figure 2. The change of the k value implies in particular that the random phase will last longer than in the previous runs as it looks for a larger set of discovered links. This induces a delay before the beginning of the second phase of the strategies which should in turn decrease their efficiency as they have less queries to test the existence of the links.

Surprisingly though, the efficiency of the strategies does not seem to be affected. The amount of discovered links after the same number of queries is for instance comparable in the *contact* network case (around 21% of the existing links). This can be explained by the fact that while searching for the 1 500 links, the random phase has improved the partial knowledge of the network topology. It is very likely then that the highly connected nodes have emerged more significantly during this phase. Thus the ordering used by the elaborated strategies, based on the *degree principle*, is in turn more pertinent.

The plots based on the *comment* and *symmetric-comment* networks also show that the behaviour of the strategies can be very similar in some cases. This suggests to investigate other criteria to sort out their efficiency.

D. Measurement bias

Until now, we focused on our ability to discover many links with as few link queries as possible. However, different strategies discover different links, which may have consequences on the properties of the obtained samples: they may be biased by the measurement strategy, and biased differently depending on the strategy we use. This can be observed visually in Figure 3 for instance, and confirmed by the statistics given in Table II.

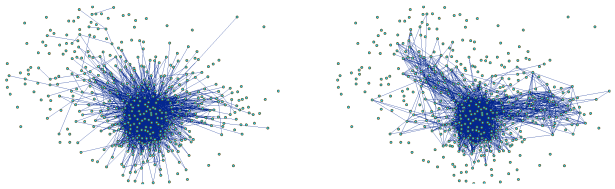

Fig. 3. Drawings of samples obtained with the *complete* (left) and *tbf-complete* (right) strategies after the 20 000 link queries. The position of the nodes is the same in the two drawings (it is obtained by a classical graph drawing algorithm ran on the actual network), which makes it possible to observe visually that the links discovered by each strategy are not the same.

	m'	δ	avg deg	max deg	cc	tr
Reference	21298	0.002	35.5	1708	0.083	0.124
random	6307	0.000	2.1	38	0.001	0.001
$\vee$*-random*	6248	0.001	3.1	123	0.133	0.120
c_{1500}	9840	0.001	13.0	1708	0.061	0.422
tbf_{1500}	2289	0.024	54.5	663	0.175	0.208
$tbfc_{1500}$	7717	0.003	20.0	1708	0.085	0.371
$\vee$-$tbfc_{1500}$	8789	0.002	17.7	1708	0.072	0.388

TABLE II
Main statistical properties (number of links finally discovered, density δ, average degree, maximal degree, clustering coefficient and transitivity ratio) of the samples obtained by each measurement strategy with $k = 1500$ applied on the *symmetric-comment* network. We also display the properties of the actual network (first row), for comparison.

These experiments clearly show that the observed properties are biased by the measurement (they are not the same as the ones of the actual network), and moreover that different strategies lead to different bias.

One can notice for instance that the *complete* and the *test-between-found* strategies induce a very different bias on

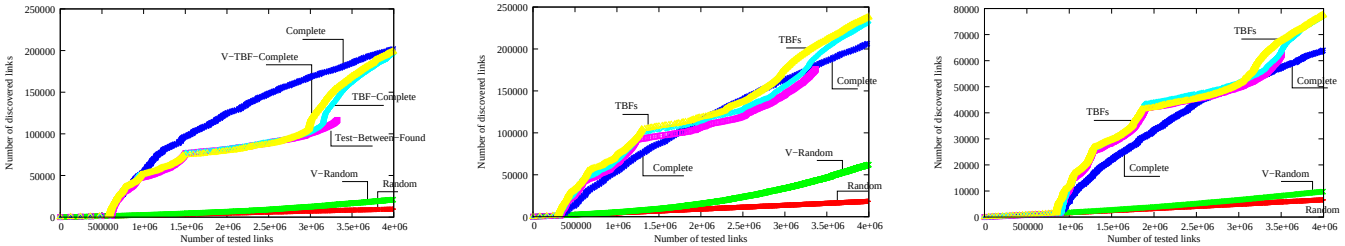

Fig. 2. Number of links (vertical axis) discovered by each strategy as a function of the number of link queries performed (horizontal axis) for each of our three graphs (from left to right: *contact*, *comment* and *symmetric-comment*), with $k = 1\,500$. The *tbf*, *tbf-complete* and $\vee$-*tbf-complete* strategies are indicated as a unique curve (named TBFs) in the two last plots as the three curves overlap eachother.

the properties. It is likely to be due to the fact that the number of involved nodes (m') is very low for the second strategy but all the possible links between them have been tested. This means in particular that all the possible triangles have been discovered which leads naturally to an over-evaluation of the clustering coefficient. The strict opposite happens in the *complete* case since many nodes are discovered but the links between them are not directly tested.

In some specific cases though, the values are correctly evaluated by the strategies. Strategy $\vee$-*random*, for instance, gives a correct value of the transitivity ratio. This is well explained by the strategy itself that tests the existence of the third link of a triangle as soon as two nodes appear to have a common neighbor.

It is also worth noticing that the mixed strategies have a better evaluation of the clustering coefficient than other strategies. This can be explained by the fact that, as the name suggests, they mix the effects of the different strategies. In particular, the over-evaluation of this property given by the *test-between-found* phase seems to be compensated by the under-evaluation of the *complete* phase.

These observations suggest to put in perspective the quantitative assessments of the runs and to try integrating the qualitative point of view in the evaluation of the efficiency of the strategies.

VI. Conclusion and perspectives

In this paper, we studied the problem of measuring large complex networks when the measurement operation consists in testing the existence of a link between two nodes. We proposed different strategies for ordering the link queries in order to minimize their number while maximizing the number of discovered links. Those strategies rely on the expected statistical properties of the network in order to predict the existence of the links and we tested this approach on several real-world networks based on the Flickr database.

The empirical results confirmed that the principles underlying the development of the strategies are relevant in this measurement context. The experiments showed that the elaborated strategies made a huge improvement compared to the random approach. But they also raised the question of accounting for the bias they induce on the extracted samples. It turned out that the different strategies gave different evaluations of the statistical properties of the original networks. This result suggests to try combining them in order to compensate those unwanted effects, which is what we plan to investigate more specifically in the future.

Going further, we also intend to extend the kind of real-world networks on which test the strategies, add measurement properties to the list of statistical properties considered (such as the assortativity and the degree-degree correlation) and try to adapt the strategies to the directed graphs.

References

[1] B. Wellman, "The network is personal: Introduction to a special issue of social networks," *Social Networks*, vol. 29, no. 3, pp. 349–356, July 2007.

[2] M. Bouvel, V. Grebinski, and G. Kucherov, "Combinatorial search on graphs motivated by bioinformatics applicatitons: a brief survey," in *Proceedings of the 31st International Workshop on Graph-Theoretic Concepts in Computer Science*, ser. Lecture Notes in Computer Science, D. Kratsch, Ed., vol. 3787. Springer Verlag, 2005, pp. 16–27.

[3] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature Genetics*, vol. 31, no. 1, pp. 60–63, May 2002.

[4] A. Lakhina, J. Byers, M. Crovella, and P. Xie, "Sampling biases in ip topology measurements," in *IEEE Infocom*, 2003.

[5] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, "On the bias of traceroute sampling," in *ACM STOC*, 2005.

[6] J.-L. Guillaume and M. Latapy, "Relevance of massively distributed explorations of the internet topology: simulation results," in *IEEE Infocom*, 2005.

[7] M. Latapy and C. Magnien, "Complex network measurements: Estimating the relevance of observed properties," in *IEEE Infocom*, 2008.

[8] D. Stutzbach, R. Rejaie, N. G. Duffield, S. Sen, and W. Willinger, "Sampling techniques for large, dynamic graphs," in *IEEE Infocom*, 2006.

[9] ——, "On unbiased sampling for unstructured peer-to-peer networks," in *IMC*, 2006.

[10] L. Dall'Asta, J. Alvarez-Hamelin, A. Barrat, A. Vazquez, and A. Vespignani, "A statistical approach to the traceroute-like exploration of networks: theory and simulations," in *CAAN*, 2004.

[11] A. Clauset, C. Moore, and M. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.

[12] D. Liben-nowell and J. Kleinberg, "The link prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[13] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, 1998.

[14] T. Schank and D. Wagner, "Finding, counting and listing all triangles in large graphs, an experimental study," in *Workshop on Experimental and Efficient Algorithms (WEA)*, 2005.

[15] ——, "Approximating clustering coefficient and transitivity," *Journal of Graph Algorithms and Applications (JGAA)*, vol. 9:2, pp. 265–275, 2005.

[16] C. Prieur, D. Cardon, J.-S. Beuscart, N. Pissard, and P. Pons, "The stength of weak cooperation: A case study on flickr," 2008, http://arxiv.org/abs/0802.2317.