Crawling Social Internetworking Systems

Francesco Buccafurri
DIMET Dept.
Univ. of Reggio Calabria
Italy
e-mail: bucca@unirc.it

Gianluca Lax
DIMET Dept.
Univ. of Reggio Calabria
Italy
e-mail: lax@unirc.it

Antonino Nocera
DIMET Dept.
Univ. of Reggio Calabria
Italy
e-mail: a.nocera@unirc.it

Domenico Ursino
DIMET Dept.
Univ. of Reggio Calabria
Italy
e-mail: ursino@unirc.it

Abstract—In new generation social networks, we expect that the paradigm of Social Internetworking Systems (SISs, for short) will be more and more important. In this new scenario, the role of Social Network Analysis is of course still crucial but the preliminary step to do is designing a good way to crawl the underlying graph. While this aspect has been deeply investigated in the field of social networks, it is an open issue when moving towards SISs. Indeed, we cannot expect that a crawling strategy which is good for social networks, is still valid in a Social Internetworking Scenario, due to its specific topological features. In this paper, we first confirm the above claim and, then, define a new crawling strategy specifically conceived for SISs. Finally, we show that it fully overcomes the drawbacks of the state-of-the-art crawling strategies.

Keywords: Social Networks, Crawling Strategies, Social Internetworking Systems

I. INTRODUCTION

In the last years, social networks have become one of the most popular communication media on the Internet [14]. The resulting universe is a constellation of several social networks, each forming a community with specific connotations, also reflecting multiple aspects of people personal life. Despite this inherent heterogeneity, the possible interaction among distinct social networks is the basis of a new emergent internetworking scenario enabling a lot of strategic applications, whose main strength will be just the integration of possibly different communities yet preserving their diversity and autonomy. This concept is very recent and only a few commercial attempts to implement Social Internetworking Systems (SISs, for short) have been proposed [1], [2], [3], [4]. In this new scenario, the role of Social Network Analysis [15], [21], [11], [25], [8], [24], [23] is of course still crucial in studying the evolution of structures, individuals, interactions, and so on, and in extracting powerful knowledge from them. But the necessary prerequisite is to have a good way to crawl the underlying graph. In the past, several crawling strategies for single social networks have been proposed. Among them the most representative ones are Breadth First Search (BFS, for short) [25], Random Walk (RW, for short) [20] and Metropolis-Hastings Random Walk (MH, for short) [13]. They were largely investigated for single social networks highlighting their pros and cons [13], [17]. But, what happens when we move towards Social Internetworking Scenarios? The question opens in fact a new issue that, to the best our knowledge, has not been investigated in the literature. Indeed, this issue is far to be trivial, since we cannot presumably expect that a crawling strategy which is good for social networks, is still valid in a Social Internetworking Scenario, due to its specific topological features.

This paper gives a contribution in this setting. In particular, through a deep experimental analysis of the above existing crawling strategies, conducted in a multi-social-network setting, it reaches the conclusion that they are little adequate to this new context, thus enforcing the need of designing new crawling strategies specific for SISs. Starting from this result, it gives its second important contribution, consisting in the definition of a new crawling strategy, called Bridge-Driven Search (BDS, for short), which relies on a feature strongly characterizing a SIS. Indeed BDS is centered on the concept of bridge, which represents the structural element that interconnects different social networks. Bridges are those nodes of the graph corresponding to users who joined more than one social network and explicitly declared their different accounts. By an experimental analysis we show that BDS fits the desired features, thus overcoming the drawbacks of existing strategies.

The plan of this paper is as follows: Section II presents related literature. The state-of-the-art crawling strategies BFS, RW and MH are analyzed in Section III, showing why they are not adequate for a SIS. In Section IV, our crawling strategy is illustrated. Section V is devoted to the presentation of the experiments carried out to evaluate BDS. Finally, Section VI provides a look at our future research efforts.

II. RELATED WORK

With the increase of both the number and the dimension of social networks the development of crawling strategies became a very challenging issue. For instance, the problem of sampling from large graphs is discussed in [19], [22]. Given a communication network, the approach of [12] aims at recognizing its overall design as well as at identifying important nodes and links in it. An investigation on the statistical properties of sampled scale-free networks is proposed in [18]. In [7], the authors compare the structures of Cyworld, MySpace and Orkut. In particular, they analyze the degree distribution, the clustering property, the degree correlation and the evolution over time of Cyworld. Some methods to produce a small realistic sample from a large real network are presented in [16]. In [25], the social network graph crawling problem is
investigated in such a way as to answer questions like: (i) how fast crawlers into consideration discover nodes/links; (ii) how different social networks and the number of protected users affect crawlers; (iii) how major graph properties are studied. A framework of parallel crawlers based on BFS and operating on eBay is described in [10]. In [17], the impact of different graph traversal techniques (namely, BFS, DFS, Forest Fire, Snowball Sampling) on the computation of the average node degree of a network is analyzed. Finally, an analysis of the Facebook friendship graph is proposed in [13].

III. Motivations

Several crawling strategies for single social networks have been proposed in the literature. Among these strategies, two very popular ones are BFS and RW. BFS implements the classical Breadth First Search visit, whereas RW selects the next node to be visited uniformly at random among the neighbors of the current node. A more recent strategy is MH [13]. At each iteration, it randomly selects a node \( w \) from the neighbors of the current node \( v \). Then, it randomly generates a number \( p \) belonging to the real interval \([0, 1]\). If \( p \leq \frac{\Gamma(v)}{\Gamma(w)} \), where \( \Gamma(v) \) (\( \Gamma(w) \), resp.) is the outdegree of \( v \) (\( w \), resp.), then it moves from \( v \) to \( w \). Otherwise, it stays in \( v \). Observe that the higher the degree of a node, the higher the probability that MH discards it.

In the past, these crawling strategies were deeply investigated when applied on a single social network. This analysis showed that none of them is always better than the other ones. Indeed, each of them could be the optimal one for a specific set of analyses. However, no investigation about the application of these strategies in a SIS has been carried out. Thus, we have no evidence that they are still valid in this new context. In order to reason about this, let us start by considering a structural peculiarity of a SIS. It is the existence of bridges, which, we recall, are those nodes of the graph corresponding to users who joined more than one social network and explicitly declared their different accounts. We expect that these nodes play a crucial role in the crawling of a SIS since they allow the crossing of different social networks, thus discovering the SIS intrinsic nature (in fact related to interconnections). Bridges are not “standard” nodes, due to their role; as a consequence, we cannot see a SIS just as a huge social network. Besides these intuitive considerations about bridges, we can help our reasoning also with two results obtained in [9], namely: Fact (i) the fraction of bridges in a social network is low, and Fact (ii) bridges have high degrees on average.

Now, the question is, what about the capability of existing crawling strategies of finding bridges? The deep knowledge about BFS, RW and MH provided by the literature allows the following conjectures to be drawn:

1) BFS tends to explore a local neighborhood of the seed it starts from. As a consequence, if bridges are not present in this neighborhood or their number is low (and this is highly probable due to Fact (ii)), the crawled sample fails in covering many social networks. Furthermore, it is well known that BFS tends to favor power users and, therefore, presents bias in some network parameters (e.g., the average degree of the nodes of the crawled portions are overestimated) [17].

2) Differently from BFS, RW does not consider only a local neighborhood of the seed. In fact, it selects the next node to be visited uniformly at random among the neighbors of the current node. Again, due to Fact (i), the probability that RW selects a bridge as the next node is low. As a consequence, the crawled sample does not cover many social networks and, if more than one social network is represented in it, the coupling degree of the crawled portions of social networks is low. Finally, analogously to BFS, RW tends to favor power users (and, consequently, to present bias in some network parameters [17]); this feature only marginally influences the capability of RW to find bridges since, in any case, their number is low.

3) MH has been conceived to unfavor power users and, more in general, nodes having high degrees which are, instead, favored by BFS and RW. It performs very well in a single social network [13] especially in the estimation of the average degree of nodes. However, due to Fact (ii), MH will penalize bridges. As a consequence, the sample crawled by MH does not cover many social networks present in the SIS.

In sum, from the above reasoning, we expect that both BFS, RW and MH are substantially inadequate in the context of SISs. As it will be described in Section V, this conclusion is fully confirmed by a deep experimental campaign, which clearly highlights the above drawbacks. Thus we need to design a specific crawling strategy for SISs.

IV. Proposed Crawling Strategy

In the design of the proposed crawling strategy we start by the analysis of some aspects limiting BFS, RW and MH in a SIS, in order to overcome them. Recall that BFS performs a breadth first search on a local neighborhood of a seed. Now, the average distance between two nodes of a single social network is generally less than the one between two nodes of different social networks. Indeed, in order to pass from a social network to another, it is necessary to cross a bridge, and since bridges are few, it could be necessary to generate a long path before reaching one of them. As a consequence, the local neighborhood considered by BFS includes one or a small number of social networks. To overcome this problem, a depth first search, instead of a breadth first search, could be performed. For this purpose, the way of proceeding of RW and MH could be included in our crawling strategy. However, since the number of bridges in a social network is low, the simple choice to go in-depth blindly does not favor the crossing from a social network to another. Even worse, since MH penalizes the nodes with a high degree, it tends to unfavor bridges, rather than favor them. Again, in the above reasoning, we have exploited Facts (i) and (ii) introduced in the previous section. A solution that overcomes the above problems could consist in implementing a “non-blind” depth first search in such a
way as to favor bridges in the choice of the next node to visit. This is the choice we have done, and the name we give to our strategy, i.e., Bridge-Driven Search (BDS, for short), clearly reflects this approach. However, in this way, it becomes impossible to explore (at least partially) the neighborhood of the current node because the visit proceeds in-depth very quickly and, furthermore, as soon as a bridge is encountered, there is a cross to another social network. The overall result of this way of proceeding would be an extremely fragmented crawled sample. In order to face this problem, given the current node, our crawling strategy explores a fraction of its neighbors before performing an in-depth search of the next node to visit.

In order to formalize our crawling strategy, we need to introduce the following parameters:

- \( \text{nf} \) (node fraction). It represents the fraction of the non-bridge neighbors of the current node which should be visited. It ranges in the real interval (0,1]. For example, when \( \text{nf} \) is equal to 1, our strategy behaves like BFS since all neighbors of the current node are selected. This parameter is used to tune the portion of the current node neighborhood which has to be taken into account and, hence, it balances the breadth and depth of the visit.

- \( \text{bf} \) (bridge fraction). It represents the fraction of the bridge neighbors of the current node which should be visited. Like \( \text{nf} \), it ranges in the real interval (0,1]. Clearly, this parameter is greater than 0 in order to allow the visit of at least one bridge (if any), thus resulting in crossing to another social network.

- \( \text{btf} \) (bridge tuning factor). It is a real number belonging to [0,1] which allows the filtering of the bridges to be visited among the available ones, on the basis of their degree. Its role will be explained in the following.

The pseudo-code of BDS is shown in Algorithm 1.

V. EXPERIMENTS

In this section, we present our experiment campaign conceived to determine the performances of BDS and to compare it with BFS, RW and MH when they operate in a SIS.

A. Techniques and Metrics

In order to perform our experiments, we implemented the four crawling strategies considered in this paper. Since we wanted to analyze the behavior of these strategies on a SIS, we had to extract not only connections among the accounts of different users in the same social network but also connections among the accounts of the same user in different social networks. In order to detect these connections, two standards encoding human relationships, namely XFN [6] and FOAF [5], are generally exploited. Interestingly enough, they allow the identification of bridge nodes as those nodes linked by more edges; suitable instructions are present in XFN and FOAF to define these edges and, ultimately, to derive bridges.

In our experiments, we considered a SIS consisting of four social networks, namely Twitter, LiveJournal, YouTube and Flickr. These networks are compliant with the XFN and FOAF standards and have been largely analyzed in the past in Social Network Analysis. For our experiments, we exploited a server equipped with a 2 Quad-Core E5440 processor and 16 GB of RAM with the CentOS 6.0 Server operating system. We performed the crawling tasks from November 5, 2011 to January 22, 2012. Collected data can be found at the URL http://www.ursino.unirc.it/bds.html.

A first needed step was to define reasonable metrics able to evaluate the performances of crawlers operating on a SIS. Even though this point could appear very critical and also prone to unfair choices, it is immediate to realize that the following chosen metrics are a good way to highlight some desired features of a crawling strategy in a SIS:

1) Bridge Ratio (BR): this is a real number in the interval [0,1] defined as the ratio of the number of the bridges discovered to the number of all the nodes in the sample.

2) Crossings (CR): this is a non-negative integer and measures how many times the crawler switches from one social network to another.

3) Covering (CV): this is a positive integer and measures how many different social networks are visited by the crawler.

4) Balancing (BL): this is a non-negative real number and is defined as the standard deviation of the percentages

---

**Algorithm 1 BDS**

Notation: We denote by \( N(x) \) the function returning the number of the non-bridge neighbors of the node \( x \), and by \( B(x) \) the function returning the number of the bridges belonging to the neighborhood of \( x \).

**Input:** \( s \): the seed node  
**Output:** \( \text{SeenNodes}, \text{VisitedNodes}: \) a set of nodes

**Algorithm: BDS**

```
Notation
Input s: the seed node
Output SeenNodes, VisitedNodes: a set of nodes
Constant \( n_{it} \) {The number of iterations}
Constant \( n_{f} \) {The node fraction parameter}
Constant \( b_{f} \) {The bridge fraction parameter}
Constant \( b_{tf} \) {The bridge tuning factor parameter}
Variable \( v, w: \) a node
Variable \( p: \) a number in the real interval (0,1)
Variable \( c: \) an integer number
Variable \( \text{NodeQueue}: \) a queue of nodes
Variable \( \text{BridgeSet}: \) a set of bridge-nodes

1: SeenNodes\(=\emptyset \), VisitedNodes\(=\emptyset \), NodeQueue\(=\emptyset \), BridgeSet\(=\emptyset \)
2: insert \( s \) into NodeQueue
3: for \( i := 1 \) to \( n_{it} \) do
4: poll the queue and extract a node \( v \)
5: insert \( v \) into VisitedNodes
6: insert all the nodes adjacent to \( v \) into SeenNodes
7: if \( (B(v) \geq 1) \) then
8: clear NodeQueue
9: \( e := 0 \)
10: while \( c < \lfloor b_{f} \cdot B(v) \rfloor \) do
11: let \( w \) be one of the bridge-nodes adjacent to \( v \) not in BridgeSet selected uniformly at random
12: generate uniformly at random a number \( p \) in the real interval (0,1)
13: if \( (p \cdot b_{tf}) \leq \frac{N(v)}{N(v) + B(v)} \) then
14: insert \( w \) into NodeQueue and BridgeSet
15: \( c := c + 1 \)
16: end if
17: end while
18: else
19: \( e := 0 \)
20: while \( c < \lfloor n_{f} \cdot N(v) \rfloor \) do
21: let \( w \) be one of the nodes adjacent to \( v \) selected uniformly at random
22: generate uniformly at random a number \( p \) in the real interval (0,1)
23: if \( (p \leq \frac{N(v)}{N(v) + B(v)}) \) then
24: insert \( w \) into NodeQueue
25: \( c := c + 1 \)
26: end if
27: end while
28: end if
29: end for
```
of nodes discovered for each social network w.r.t. the overall number of nodes discovered in the sample. Observe that Balancing ranges from 0 to a maximum value (for instance, 50 in case of 4 social networks), corresponding to the case in which all sampled nodes belong to a social network.

5) **Degree Bias (DB):** this is a real number computed as the root mean squared error, for each social network of the SIS, of the average node degree estimated by the crawler and the one estimated by MH, which is considered the best one in estimating the node degree for a social network in the literature [17], [13]. If the crawled sample does not cover one or more social networks, then these last ones are not considered in the computation of Degree Bias.

It is worth pointing out that, as for the first three metrics, the higher their value the higher the performance of the crawling strategy. By contrast, as for the fourth and the fifth metrics, the lower their values the higher the performance of the crawling strategy. Observe that Covering is related to the crawler capability of covering many social networks. Balancing is related to the crawler capability of uniformly sampling all the social networks. Furthermore, observe that, even though one could intuitively think that a fair sampling should sample different social networks proportionally to their respective overall size, a similar behavior of the crawler could result in incomplete samples in case of high variance of these sizes. Indeed, it could happen that some small social networks would be not represented in the sample or represented in an insufficient way. Bridge Ratio and Crossing are related to the coupling degree, while Degree Bias to the average degree. Finally, we note that the defined metrics are not completely independent from each other. For instance, if $BR = 0$, then $CR$ and $CV$ are also 0. Analogously, the value of $CR$ influences both $CV$ and $BL$.

### B. Analysis of BFS, RW, MH and BDS

In this section, we analyze the performances of BFS, RW, MH and BDS, when applied on a SIS.

By considering the three classical crawling strategies, namely MH, BFS and RW, BDS is examined later. As for the first three strategies, we can draw the following conclusions:

1) The value of Bridge Ratio is very low for all these crawling strategies. For MH there are 2.5 bridges for each 1000 crawled nodes on average. BFS behaves worse than MH and RW is the worst one.

2) The value of Crossing is generally low for all these crawling strategies. For BFS an increase of the value can be observed only when the seed belongs to YouTube. RW shows again the worst value. This result is clearly related to the low value of Bridge Ratio, since the few discovered bridges do not allow the crawlers to sufficiently cross different social networks.

3) The value of Covering is quite low for all these techniques. On average only two of the social networks of the SIS are visited. Also this result is related to the low values of Bridge Ratio and Crossing.

4) The value of Balancing is very high for all these techniques, very close to the maximum one (i.e., 50). This indicates that, as far as this metric is concerned, they behave very badly. Indeed, it happens that they often stay substantially bounded in the social network of the starting seed. BFS and RW are even worse than MH.

5) As for the average degrees of nodes, it is well known that MH is the crawling strategy that best estimates them in a single social network [17], [13]. From the previous reasoning on Balancing it is possible to deduce that each run of MH visits one or at most two social networks. As a consequence, we can assume that the average degrees it provides are also those of reference for the social networks of the SIS, provided that at least one run of MH starting from each social network is performed.
On the basis of these reference values, we detect that BFS presents a high value of Degree Bias. This fact is well known in the literature for a scenario consisting of a single social network, and thus we confirm this conclusion also in the context of SISs. RW is even worse than BFS.

In sum, we may conclude that the conjectures given in Section III are fully confirmed by experiments. Now we have to see how our crawling strategy behaves on the same SIS.

In order to analyze BDS we first performed a large set of experiments to identify the best values of $n_f$, $b_f$ and $b_{tf}$. Due to space reasons, we cannot report here the details of these experiments. The interested reader can find them at the URL http://www.ursino.unirc.it/bds.html. We only say that a parameter configuration generally guaranteeing satisfactory values for all the metrics into consideration is $n_f = 0.10$, $b_f = 0.25$, $b_{tf} = 0.25$. However, there could be some applications in which some metrics are much more important than the other ones. BDS is highly flexible and, in these cases, allows the choice of the configuration which favors that metric. The values of our metrics obtained by BDS with the chosen configuration are reported in the last part of Table I. From the analysis of these values it clearly emerges that, when operating on a SIS, BDS highly outperforms the other approaches. The only exception is MH for Degree Bias since, according to [13], [17], we have assumed that MH is the best method to estimate the average node degree. However, also for this metric, BDS obtains very satisfactory results. As a final remark we highlight that, besides the capability shown by BDS of crossing through different social networks, overcoming the drawbacks of compared crawler strategies, BDS presents a good behavior also from an intra-social-network point of view. This claim is supported from both the results obtained for the Degree Bias, and the consideration that our crawling algorithm, in absence of bridges, can be located between BFS and MH, thus producing intra-social-network results that reasonably cannot differ significantly from the ones of these strategies.

VI. CONCLUSION AND FUTURE WORK

We think that SIS analysis is a very promising research field and so we plan to perform further research efforts in the future. In particular, one of the most challenging issue could be the improvement of the BDS crawling strategy in such a way that the values of $n_f$, $b_f$ and $b_{tf}$ dynamically change during the crawling activity to adapt to the specificities of the crawled SIS. Moreover, we plan to better investigate the properties of BDS also as single-social-network crawler, also to help the setting of the above parameters. Finally, we plan to investigate the possible connections of our approach with the information integration ones as well as to deal with the privacy issue in crawling with SIS settings.

Acknowledgment

This work was partially funded by the Italian Ministry of Research through the PRIN Project EASE (Entity Aware Search Engines).