Experiences Using BDS: A Crawler for Social Internetworking Scenarios

Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino

Abstract In new generation social networks, we expect that the paradigm of Social Internetworking Scenarios (SISs) 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, good for social networks, is still valid in a Social Internetworking scenario, due to the specific topological features of this scenario. In this paper, we first confirm the above claim and then, define a new crawling strategy specifically conceived for SISs, which overcomes the drawbacks of the state-of-the-art crawling strategies. After this, we exploit this crawling strategy to investigate SISs to understand their main properties and features of their main actors (i.e., bridges).

1 Introduction

In recent years, (online) social networks (OSN, for short) have become one of the most popular communication media on the Internet [31]. 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

F. Buccafurri (✉) • G. Lax • A. Nocera • D. Ursino
DIIES, University Mediterranea of Reggio Calabria Via Graziella,
Località Feo di Vito, 89122 Reggio Calabria, Italy
e-mail: bucca@unirc.it; lax@unirc.it; a.nocera@unirc.it; ursino@unirc.it

Ş. Gündüz-Öğüdücü and A.Ş. Etaner-Uyar (eds.), *Social Networks: Analysis and Case Studies*, Lecture Notes in Social Networks, DOI 10.1007/978-3-7091-1797-2_8,

Scenarios (SISs, for short) have been proposed [9, 10, 21, 22, 24, 51]. In this new scenario, the role of Social Network Analysis [4, 15, 34, 44, 53, 57, 62] is of course still crucial in studying the evolution of structures, individuals, interactions, and so on, and in extracting powerful knowledge from them. But an important 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) [62], Random Walk (RW, for short) [41] and Metropolis-Hastings Random Walk (MH, for short) [27]. They were largely investigated for single social networks highlighting their pros and cons [27, 36]. But, what happens when we move towards Social Internetworking Scenarios? In fact, the question opens a new issue that, to the best our knowledge, has not been investigated in the literature. Indeed, this issue is far from being trivial, because we cannot expect that a crawling strategy, good for social networks, is still valid in a Social Internetworking Scenario, due to the specific topological features of this scenario.

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, enforcing the need of designing new crawling strategies specific for SISs. Starting from this result, this paper gives a 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, overcoming the drawbacks of existing strategies.

As a third important contribution, with the support of such a crawler specifically designed for SISs, we extract data from SISs to detect the main properties of this new kind of scenario and, especially of its main actors, which are bridges. The analysis of bridges, aiming at estimating both classical Social Network Analysis parameters and new specific ones, is conducted in such a way as to discover the nature of bridges in a very deep fashion. For this purpose, a large number of experiments is performed, to derive knowledge about the following topics:

- distribution of the contact number of bridges (hereafter, bridge degree) and non-bridges;
- correlation between bridges and power users (i.e., nodes having a very high degree, generally higher than the average degree of the social network joined by them);
- existence of preferential ties among bridges;
- centrality of bridges in a SIS and in its single social networks.

The results of our analysis provide knowledge about these topics with a strong experimental support and discover even unexpected conclusions about bridges and,

in general, a complete knowledge of these crucial elements of Social Internetworking Scenarios.

The plan of this paper is as follows: in Sect. 2, we present related literature. In Sect. 3, we illustrate and validate our Bridge Driven Search approach. In Sect. 4, we describe our experiences devoted to define the main features of SISs and of bridges. Finally, in Sect. 5, we draw our conclusions and we present possible future issues in this research field.

2 Related Literature

In this section, we survey the scientific literature related to our paper. In particular, we first describe the most known techniques proposed to crawl social networks and then we focus on the approaches proposed for Social Network Analysis.

Concerning the former issue, we observe that with the increase in both the number and the dimension of social networks, the development of approaches to sample social networks has become a very challenging issue. The problem of sampling from large graphs is discussed in [38]. In this paper, the authors aim at answering questions such as: (1) which sampling method to use; (2) how small can the sample size be; (3) how to scale up the measurements of the sample to get estimates for larger graphs; (4) how success can be measured. In their activity they consider several sampling methods and check the goodness of their sampling strategies on several datasets.

A technique based on both sampling and the randomized notion of *focus* is proposed in [52]. This method stores samples in a relational database and favors the visualization of massive networks. In this work, the authors specify features frequently characterizing massive networks and analyze the conditions allowing their preservation during the sampling task. An investigation of the statistical properties of sampled scale-free networks is proposed in [37]. In this paper, the authors present three sampling methods, analyze the topological properties of obtained samples, and compare them with those of the original network. Furthermore, they explain the reasons of some emerged biased estimations and provide suitable criteria to counterbalance them.

Methods to produce a small realistic sample from a large real network are presented in [33]. Here, the authors show that some of the proposed methods maintain the key properties of the initial graph even with a sample size down to 30 %. In [62], the social network graph crawling problem is investigated in such a way as to answer questions such as: (1) how fast crawlers into consideration discover nodes/links; (2) how different social networks and the number of protected users affect crawlers; (3) how major graph properties are studied. All these investigations are performed by analyzing samples derived from four social networks, i.e. Flickr, LiveJournal, Orkut and YouTube.

A framework of parallel crawlers based on BFS and operating on eBay is described in [14]. This framework exploits a centralized queue. The crawlers operate

independently from each other so that the failure of one of them does not influence the others. In spite of this, no redundant crawling occurs. In [36], the impact of different graph traversal techniques (namely, BFS, DFS, Forest Fire and Snowball Sampling) on the computation of the average node degree of a network is analyzed. In particular, the authors quantify the bias of BFS in estimating the node degree w.r.t. the fraction of sampled nodes. Furthermore, they show how this bias can be corrected. An analysis of the Facebook friendship graph is proposed in [27]. In this activity, the authors examine and compare several candidate crawling strategies, namely BFS, Random Walk, Metropolis-Hastings Random Walk and Re-Weighted Random Walk. They investigate also diagnostics to assess the quality of the samples obtained during the data collection process.

Concerning the main difference between the new crawler BDS proposed in our paper and the above crawling techniques, we highlight the fundamental difference is that the above techniques are not specifically designed to operate effectively on a SIS. This will be confirmed by the experimental analysis provided in Sect. 3.2.

As far as the latter issue dealt with in this section (i.e., Social Network Analysis) is concerned, we observe that studies on Social Networks attracted mainly sociologists. For instance, [58] introduced the six-degrees of separation and the small-world theories. The effects of these theories are analyzed in [19]. Granovetter [28] showed that a Social Network can be partitioned into "strong" and "weak" ties, and that strong ties are tightly clustered. In a second time, with the development of OSNs, Social Network Analysis attracted computer scientists and many studies have been proposed, which investigate the features of one OSN or compare more OSNs. Most of them collect data from one or more OSNs, map these data onto graphs and analyze their structural properties. These approaches are based on the observation that topological properties of graphs may be reliable indicators of the behaviors of the corresponding users [31].

Studies about how an attacker discovers a social graph can be found in [7, 32]. The sole purpose of the attacker is to maximize the number of nodes/links that can be discovered. As a consequence, these two papers do not examine other issues, such as biases.

In [2], 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. After this, they use Cyworld to evaluate the snowball sampling method exploited to sample MySpace and Orkut. Finally, they perform several interesting analyses on the three social networks.

Given a communication network, the approach of [26] aims at recognizing the network topology and at identifying important nodes and links in it. Furthermore, it proposes several compression schemes exploiting auxiliary and purely topological information. Finally, it examines the properties of such schemes and analyzes what structural graph properties they preserve when applied to both synthetic and real-world networks.

In [43], the authors present a deep investigation of the structure of multiple OSNs. For this purpose, they examine data derived from four popular OSNs, namely Flickr, YouTube, LiveJournal and Orkut. Crawled data regard publicly accessible user links

on each site. Obtained results confirm the power law, small-world and scale-free properties of OSNs and show that these contain a densely connected core of high-degree nodes.

In [35], the authors focus on analyzing the giant component of a graph. Moreover, they define a generative model to describe the evolution of the network. Finally, they introduce techniques to verify the reliability of this model. In [3], the authors investigate the main features of groups in LiveJournal and propose models that represent the growth of user groups over time. In [40], data crawled from LiveJournal are examined to investigate the possible correlations between friendship and geographic location in OSNs. Moreover, the authors show that this correlation is strong. Carrington et al. [12] proposes a methodology to discover possible aggregations of nodes covering specific positions in a graph (e.g., central nodes), as well as very relevant clusters. Still on clustering, De Meo et al. [18] recently proposed an efficient community detection algorithm, particularly suited for OSNs, and tested its performance against a large sample of Facebook (among other OSN samples), observing the emergence of a strong community structure. In [50], the authors propose *Social Action*, a system based on attribute ranking and coordinated views to help users to systematically examine numerous Social Network Analysis measures. In [13], the authors present an analysis of Facebook devoted to investigate the friendship relationships in this OSN. To this purpose, they examine the topological properties of graphs representing data crawled from this OSN by exploiting two crawling strategies, namely BFS and Uniform Sampling. A further analysis of Facebook can be found in [59]. In this paper, the authors crawled Facebook by means of BFS and formalized some properties such as assortativity and interaction. These can be verified in small regions but cannot be generalized to the whole graph.

Monclar et al. [45], Ghosh and Lerman [23], Onnela and Reed-Tsochas [49], and Romero et al. [54] present approaches for the identification of influential users, i.e. users capable of stimulating others to join OSN activities and/or to actively operate in them. In [1, 39, 55], the authors suitably model the blogosphere to perform leader identification. In [42], the authors first introduce the concept of starters (i.e., users who generate information that catches the interest of fellow users/readers) and, then, adopt a Random Walk technique to find starters. The authors of [47] analyze the main properties of the nodes within a single OSN that connect the peripheral nodes and the peripheral groups with the rest of the network. The authors call these nodes bridging nodes or, simply, bridges. Clearly, here, the term "bridge" is used with a meaning totally different from that adopted in our paper. The authors base their analysis on the study of the theoretical properties of their model. In [25], the authors propose a predictive model that maps social media data to tie strength. This model is built on a dataset of social media ties and is capable of distinguishing between strong and weak ties with a high accuracy. Moreover, the authors illustrate how tie strength modeling can improve social media design elements, such as privacy controls, message routing, friend introductions and information prioritization. The authors of [60] present a model for predicting the closeness of professional and personal relationships of OSN users on the basis of their behavior in the OSNs

joined by them. In particular, they analyze how the behavior of users on an OSN reflects the strength of their relationships with other users w.r.t. several factors, such as profile commenting and mutual connections.

A preliminary study about SISs and bridges has been done in [11]. However, it has been carried out by investigating samples extracted through classical crawling techniques. Berlingerio et al. [5, 6], Dai et al. [17], Mucha et al. [46], and Kazienko et al. [30] present approaches in the field of multidimensional networks. These networks can be seen as a specific case of a SIS in which each social network is specific for one kind of relationship and social networks strongly overlap. Multidimensional social networks are known as multislice networks in the literature [46].

Concerning the originality of our paper w.r.t. the above literature, we note that none of the above studies analyzes the main features of SIS. By contrast, in our paper, we provide a deep analysis on bridges, which are the key concept of a SIS.

3 The Bridge Driven Search Crawler

As pointed out in the introduction, the first main purpose of this paper is to investigate crawling strategies for a SIS. These must be able to extract not only connections among the accounts of different users in the same social network but also interconnections among the accounts of the same user in different social networks. Several crawling strategies for single social networks have been proposed in the literature. Among these strategies, two very popular ones are BFS [62] and RW [41]. The former implements the classical Breadth First Search visit, the latter selects the next node to be visited uniformly at random among the neighbors of the current node. A more recent strategy is MH [27]. 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. The pseudocode of this algorithm is shown in Algorithm 1. 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 others. Indeed, each of them can 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. To reason about this, let us start by considering a structural peculiarity of a SIS, i.e. 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 as they allow the crossing of different social networks, discovering the SIS intrinsic nature (related to interconnections). Bridges are not "standard" nodes, due to their role; thus, 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 [11], for which: *Fact (i)* the fraction of bridges in a social network

Algorithm 1: MH

Notation We denote by $\Gamma(x)$ the outdegree of the node x
Input s: a seed node
Output *SeenNodes*, *VisitedNodes*: a set of nodes
Constant n_{it} {The number of iterations}
Variable v, w: a node
Variable p: a number in the real interval (0,1)
1: *SeenNodes*:=∅, *VisitedNodes*:=∅
2: insert s into *SeenNodes* and *VisitedNodes*
3: insert all the nodes adjacent to s into *SeenNodes*
4: $v = s$
5: **for** $i := 1$ to n_{it} **do**
6: let w be one of the nodes adjacent to v selected uniformly at random
7: generate uniformly at random a number p in the real interval (0,1)
8: **if** $(p \leq \frac{\Gamma(v)}{\Gamma(w)})$ **then**
9: $v = w$
10: insert w into *VisitedNodes*
11: insert all the nodes adjacent to w into *SeenNodes*
12: **end if**
13: **end for**

is low, and *Fact (ii)* bridges have high degrees on average. This is confirmed by the experimental results presented in Table 8.

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 us to draw the following conjectures:

- 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 *(i)*), 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 [36]).
- 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 [36]. This feature only marginally influences the capability of RW to find bridges because, in any case, their number is very low.
- 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 [27] especially in the estimation of the average degree of nodes. However, due to Fact *(ii)*, it 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 Sect. 3.2, 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. This is a matter of the next section.

3.1 BDS Crawling Strategy

In the design of our new crawling strategy, we start from the analysis of some aspects limiting BFS, RW, and MH in a SIS, 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, to pass from a social network to another, it is necessary to cross a bridge, and as bridges are few, it may 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, can be done. For this purpose, the way of proceeding of RW and MH may be included in our crawling strategy. However, because 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, because MH penalizes the nodes with a high degree, it tends to unfavor bridges, rather than to 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 consists 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 do, 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 is an extremely fragmented crawled sample. To address 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.

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

- *nf* (*node fraction*). It represents the fraction of the *non-bridge* neighbors of the current node that should be visited. It ranges in the real interval (0,1]. For example, when *nf* is equal to 1, our strategy selects all the neighbors of the current node except the bridge ones. This parameter is used to tune the portion of the current node neighborhood that has to be taken into account and, hence, it balances the breadth and depth of the visit.

Algorithm 2: 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 *SeenNodes*, *VisitedNodes*: a set of nodes
Constant n_{it} {The number of iterations}
Constant nf {The *node fraction* parameter}
Constant bf {The *bridge fraction* parameter}
Constant btf {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 *NodeQueue*: a queue of nodes
Variable *BridgeSet*: a set of bridge nodes

```
1:  SeenNodes:=∅, VisitedNodes:=∅, NodeQueue:=∅, BridgeSet:=∅
2:  insert s into NodeQueue
3:  for i := 1 to n_it do
4:      poll NodeQueue and extract a node v
5:      insert v into VisitedNodes
6:      insert all the nodes adjacent to v into SeenNodes
7:      if (B(v) ≥ 1) then
8:          clear NodeQueue
9:          c = 0
10:         while (c < ⌈bf · B(v)⌉) 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 · btf ≤ N(v)/N(w)) then
14:                 insert w into NodeQueue and BridgeSet
15:                 c = c + 1
16:             end if
17:         end while
18:     else
19:         c = 0
20:         while (c < ⌈nf · N(v)⌉) 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 ≤ N(v)/N(w)) then
24:                 insert w into NodeQueue
25:                 c = c + 1
26:             end if
27:         end while
28:     end if
29: end for
```

- *bf* (*bridge fraction*). It represents the fraction of the bridge neighbors of the current node that should be visited. Like *nf*, it ranges in the real interval (0,1]. Clearly, this parameter is greater than 0 to allow the visit of at least one bridge (if any), resulting in crossing to another social network.
- *btf* (*bridge tuning factor*). It is a real number belonging to [0,1] that allows the filtering of the bridges to be visited among the available ones, on the basis of their degree. Its role will be better explained in the following.

For instance, in a configuration with $nf = 0.10$ and $bf = 0.25$, our strategy visits 10 % of the non-bridge neighbors of the current node and 25 % of the bridge neighbors of the current node.

We are now able to formalize our crawling strategy. Its pseudo-code is shown in Algorithm 2.

The algorithm exploits two data structures: a queue *NodeQueue* of nodes and a set *BridgeSet* of bridges. The former contains the nodes detected during the crawling task and that should be visited later; the latter contains the bridges that have been already met during the visit. BDS starts its visit from a seed node s that is added into *NodeQueue*. At each iteration, a new node v is extracted from *NodeQueue*; v is inserted into *VisitedNodes*, whereas all the nodes adjacent to v are put into *SeenNodes* (Lines 4–6). After this, if v has at least one bridge as neighbor, then the visit proceeds towards one or more of the bridge neighbors of v, thus switching the current social network. For this reason, *NodeQueue* is cleared (Line 8). This is necessary because, if the next nodes are polled from *NodeQueue*, the visit is brought back to the old social network improperly.

Now, in Line 10, the algorithm computes how many bridges must be selected on the basis of its setting. After this, each of these bridges, say w, is selected uniformly at random among those not previously met in the visit; w is added into *NodeQueue* and into *BridgeSet* if and only if the ratio between the v's outdegree and w's outdegree is greater than or equal to $p \cdot btf$, where p is a real random number in [0,1] (Line 13). Observe that this condition is similar to that adopted by MH to drive the selection of nodes on the basis of their degrees. In particular, when $btf = 1$, this condition coincides with the one of MH, disadvantaging high-degree bridges; when $btf = 0$, no filtering on the bridge degree is done. Clearly, values ranging from 0 to 1 result in an intermediate behavior. If no bridge has been discovered, then $\lceil nf \cdot N(v) \rceil$ non-bridges adjacent to v are randomly selected (Lines 20 and 21). Such nodes are selected according to the policy of MH. They are added into *NodeQueue*. The algorithm terminates after n_{it} iterations.

As for Lines 20–27, it is worth pointing out that, differently from MH, which selects only one neighbor for each node (thus, performing an in-depth visit), BDS has also a component (i.e., nf), which allows it to select more than just one neighbor in such a way as to make it able to take the neighborhood of the current node into account. In this way, it solves one of the problems of RW and MH discussed above.

3.2 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. As 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. To encode these connections, two standards encoding human relationships are generally exploited. The former is XFN (XHTML Friends Network) [61]. XFN simply uses an attribute, called `rel`, to specify the kind of relationship between two users. Possible values of `rel` are `me`, `friend`, `contact`, `co-worker`, `parent`, and so on. A (presumably) more complex alternative to XFN is FOAF (Friend-Of-A-Friend) [8]. A FOAF profile is essentially

an XML file describing people, their links to other people and their links to created objects. The technicalities concerning these two standards have not to be handled manually by the user. As a matter of fact, each social network has suitable mechanisms to automatically manage them in a way transparent to users, who have simply to specify their relationships in a friendly fashion.

In our experiments, we consider a SIS consisting of four social networks, namely Twitter, LiveJournal, YouTube and Flickr. They are compliant with the XFN and FOAF standards and have been largely analyzed in Social Network Analysis in the past [15, 34, 44, 62]. We argue that the relatively small number of involved social networks, as a first investigation, is adequate, expecting that the more this number, the higher the gap between standard and specific crawling strategies.

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. Collected data can be found at the URL http://www.ursino.unirc.it/ebsnam.html. (The password to open the archive is "84593453".)

3.2.1 Metrics

A first needed step was to define reasonable metrics able to evaluate the performances of crawlers operating on a SIS. Even though this point may appear very critical and prone to unfair choices, it is immediate to realize that the following chosen metrics are a good way to highlight the desired features of a crawling strategy operating 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. *Unbalancing (UB)*: this is a non-negative real number and is defined as the standard deviation of the percentages of nodes discovered for each social network w.r.t. the overall number of nodes discovered in the sample. Observe that Unbalancing ranges from 0, corresponding to the case in which each social network is sampled with an equal number of nodes, 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. For example, in a SIS consisting of four social networks, if the overall discovered nodes are 100 and the number of nodes belonging to each of the four social networks is 40, 11, 30, and 19, resp., then *UB* is equal to 12.68.
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 that estimated by MH, which is considered the best one in

estimating the node degree for a social network in the literature [27, 36]. If the crawled sample does not cover one or more social networks, then these are not considered in the computation of the Degree Bias.

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 metric, 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. Unbalancing measures the crawler capability of uniformly sampling all the social networks. Furthermore, observe that, even though one may intuitively think that a fair sampling should sample different social networks proportionally to their respective overall size, a similar behavior of the crawler results in incomplete samples in case of high variance of these sizes. Indeed, it may happen that small social networks are 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 UB.

Besides the evaluation of the crawling strategies on each of the above metrics, separately considered, it is certainly important to define a synthetic measure capable of capturing a sort of "overall" behavior of the strategies, possibly modulating the importance of each metric. A reasonable way to do this is to compute a linear combination of the five metrics, in which the coefficients reflect the importance associated with them. We call *Average Crawling Quality* (*ACQ*) this measure and define it as:

$$ACQ = w_{BR} \cdot \frac{BR}{BR_{max}} + w_{CR} \cdot \frac{CR}{CR_{max}} + w_{CV} \cdot \frac{CV}{CV_{max}} + w_{UB} \cdot (1 - \frac{UB}{UB_{max}}) + w_{DB} \cdot (1 - \frac{DB}{DB_{max}})$$

where BR_{max} (CR_{max}, CV_{max}, UB_{max}, DB_{max}, resp.) are upper bounds of Bridge Ratio (Crossings, Covering, Unbalancing, Degree Bias, resp.) that, in a comparative experiment, can be set to the maximum value obtained by the compared techniques, whereas $w_{BR}, w_{CR}, w_{CV}, w_{UB}$, and w_{DB} are positive real numbers belonging to $[0, 1]$ such that $w_{BR} + w_{CR} + w_{CV} + w_{UB} + w_{DB} = 1$. Below, we deal with the problem of setting the values of these parameters.

3.2.2 Analysis of BFS, RW and MH

In this section, we analyze the performances of BFS, RW and MH, when applied on a SIS. For this purpose, we randomly chose four seeds, each belonging to one of the social networks of our SIS, and, for each crawling strategy, we run the corresponding crawler one time for each of the four seeds. The number of iterations of each crawling run was 5,000. The overall numbers of seen nodes returned by MH,

Table 1 Performances of MH, BFS and RW

		Twitter	YouTube	Flickr	LiveJournal	Overall
MH	Bridge Ratio	0.0028	0.0038	0.0	0.0024	0.0025
	Crossing	5	3	0	4	3
	Covering	2	2	1	3	2
	Unbalancing	32.5503	40.6103	50	31.1253	38.5715
	Degree Bias					0
	Twitter Avg. Deg.	*37.9189*	*37.0789*	–	*38.2842*	*37.76067*
	YouTube Avg. Deg.	*9.6064*	*10.0379*	–	*10.4144*	*10.01957*
	Flickr Avg. Deg.	–	–	*104.3497*	–	*104.3497*
	LiveJournal Avg. Deg.	–	–	–	*40.5604*	*40.5604*
BFS	Bridge Ratio	0.0008	0.0054	0.0	0.0	0.0016
	Crossing	7	43	0	0	12.5
	Covering	2	3	1	1	1.75
	Unbalancing	49.9350	42.0286	50	50	47.9909
	Degree Bias					103.9339
	Twitter Avg. Deg.	*21.92*	–	–	–	*21.92*
	YouTube Avg. Deg.	–	*9.5106*	–	–	*9.5106*
	Flickr Avg. Deg.	–	–	*311.6118*	–	*311.6118*
	LiveJournal Avg. Deg.	–	–	–	*40.0136*	*40.0136*
RW	Bridge Ratio	0.0	0.00067	0.0	0.0	0.00017
	Crossing	0	4	0	0	1
	Covering	1	3	1	1	1.5
	Unbalancing	50	40.1363	50	50	47.5341
	Degree Bias					180.8260
	Twitter Avg. Deg.	*39.0*	*41.9489*	–	–	*40.4745*
	YouTube Avg. Deg.	–	*12.5081*	–	–	*12.5081*
	Flickr Avg. Deg.		*476.6446*	*455.3002*	–	*465.9724*
	LiveJournal Avg. Deg.	–	–	–	*43.3333*	*43.3333*

RW and BFS were 135,163, 941,303 and 726,743, respectively. The high variance of these numbers is not surprising because it is intrinsic in the way of proceeding of these algorithms.

In Table 1, we show the values of our metrics, along with the values of the other parameters we consider particularly significant (i.e., the average degree of the nodes of each social network), obtained for the four runs of MH, BFS and RW, respectively.

From the analysis of this table, we can draw the following conclusions:

- The value of *BR* is very low for all the crawling strategies. For MH there are on average 2.5 bridges for each 1,000 crawled nodes. BFS behaves worse than MH, and RW is the worst one. This behavior can be explained by the theoretical observations about BFS, MH and RW provided in Sect. 3.
- The value of *CR* is generally low for all the crawling strategies. RW shows again the worst value. This result is clearly related to the low value of *BR*, because the

few discovered bridges do not allow the crawlers to sufficiently cross different social networks.
- The value of *CV* is quite low for all the strategies. On average only two of the social networks of the SIS are visited. Also this result is related to the low values of *BR* and *CR*.
- Even though *BR*, *CR* and *CV* are generally low for all social networks, this trend is mitigated for YouTube when BFS and RW are adopted. This can be explained by the fact that the central concept in YouTube is *channel*, rather than profile. A channel has generally associated the links with the profiles of the corresponding owners present in the other social networks. As a consequence, YouTube tends to behave as a "hub" among the other social networks. This implies that the number of bridges in YouTube is higher than in the other social networks; in its turn, this implies an increase in *BR*, *CR* and *CV*. This trend is not observed for MH because this crawling technique tends to unfavor high-degree nodes, and often bridges have this characteristic (see *Fact (ii)* in Sect. 3).
- The value of *UB* is very high for all the strategies, 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. This result can be explained (*i*) by the fact that *UB* is influenced by *CR*, which, in turn, is influenced by *BR*, and (*ii*) by the previous conclusions about *BR* and *CR*.
- 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 [27, 36]. From the analysis of Table 1, we observe that when MH starts from a seed it generally stays for many iterations in the corresponding social network (this is witnessed by the high values of *UB*). As a consequence, we can assume that the average degrees are 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. This conclusion is further enforced by observing that (as shown in Table 1) MH is capable of estimating the average degree of nodes even for social networks different from that of the seed (whose number of nodes in the sample is quite low). Basing on these reference values, we detect that BFS presents a high value of *DB*. This is well known in the literature for a scenario consisting of a single social network [36], and we confirm this conclusion also in the context of SISs. The performance of RW is even worse than that of BFS.

In sum, we may conclude that the conjectures given above about the unsuitability of BFS, MH and RW to operate on a SIS are fully confirmed by our experiments. Now we have to see how our crawling strategy performs in this scenario. This is the matter of the next section.

Table 2 Performances of BDS for different values of *nf*

$bf = 0.25$, $btf = 0.25$	$nf = 0.02$	$nf = 0.10$	$nf = 0.25$	$nf = 0.50$
Bridge Ratio	0.0488	0.0965	0.1516	0.0814
Crossing	28	286	639	410
Covering	3	4	4	4
Unbalancing	26.1299	13.5103	12.2921	42.2080
Degree Bias	N.A.	19.0145	68.7413	123.705
Twitter Avg. Deg.	*40.5714*	*42.0383*	*41.0812*	*42.1161*
YouTube Avg. Deg.	*12.079*	*11.9278*	*12.8508*	*14.3556*
Flickr Avg. Deg.	*240.8333*	*129.5833*	*153.6764*	*331.5379*
LiveJournal Avg. Deg.	N.A.	*68.6234*	*168.8152*	*138.3333*

3.2.3 Analysis of BDS

To analyze BDS we performed a large set of experiments. In this section, we present the most significant ones to evaluate the impact of *nf*, *bf* and *btf*. In the configurations considered in these experiments we performed 5,000 iterations. The number of obtained seen nodes ranges from 15,585 to 473,122.

Impact of *nf*

We first evaluate the role of *nf* on the behavior of BDS. For this purpose we have fixed the other two parameters *bf* and *btf* to 0.25, and we have assigned to *nf* the following values: 0.02, 0.10, 0.25 and 0.50 (the reasons underlying the choice of discarding lower or higher values will be clear below). The results of this experiment are shown in Table 2. From the analysis of this table we observe that very low values of *nf* (i.e., *nf* about 0.02) lead to a significant decrease in *BR* and *CR*. Furthermore not all the social networks of the SIS are sampled. This behavior can be explained by the fact that, when *nf* is very low, BDS behaves as RW. This has a very negative influence on *UB*, because *BR* and *CR* influence *UB*, and does not allow the computation of *DB*, because not all social networks of the SIS are covered. As a consequence, we have decided not to report values of *nf* lower than 0.02.

By contrast, for high values of *nf* (i.e., *nf* about 0.50) we observe that *BR*, *CR* and *CV* show satisfying values. However, in this case, we obtain the worst values of *UB* and *DB*, registering for these metrics a behavior of BDS similar to that of BFS (as a matter of fact, $nf = 0.50$ implies that 50 % of the non-bridge neighbors of each node are visited). The high *UB* is explained by the fact that, even though the visit involved all the four social networks, this did not happen in a uniform fashion and some social networks have been sampled much more than the others. As for *DB*, in this case, BDS shows a behavior even worse than BFS because the presence of a high number of bridges in the sample causes the increase in the estimated average degree of the social networks (recall Fact *(ii)* introduced in Sect. 3). For this reason we do not report values of *nf* higher than 0.50.

Table 3 Performances of BDS for different values of *bf*

$nf = 0.10$, $btf = 0.25$	$bf = 0.25$	$bf = 0.50$	$bf = 0.75$	$bf = 1.00$
Bridge Ratio	0.0965	0.0875	0.0815	0.0824
Crossing	286	264	270	266
Covering	4	4	4	4
Unbalancing	13.5103	18.4622	8.4756	16.1804
Degree Bias	19.0145	21.8163	53.9505	50.8895
Twitter Avg. Deg.	*42.0383*	*40.76*	*40.4362*	*40.3348*
YouTube Avg. Deg.	*11.9278*	*12.2291*	*11.9679*	*11.7807*
Flickr Avg. Deg.	*129.5833*	*147.428*	*157.427*	*156.6471*
LiveJournal Avg. Deg.	*68.6234*	*46.4071*	*134.4459*	*127.82*

The reasoning above suggests that, to cover all the social networks of the SIS, *nf* should be higher than 0.02. However, to obtain acceptable *DB* values, it should be lower than 0.50. For this reason, we decided to fix *nf* to the intermediate value 0.10 in the study of *bf* and *btf*. In fact, this value shows a good tradeoff w.r.t. all considered metrics.

Impact of *bf*

To evaluate the impact of *bf* on the behavior of BDS we fixed *nf* to 0.10 and *btf* to 0.25. We assigned to *bf* the following values: 0.25, 0.50, 0.75 and 1. We set 0.25 as the lower bound for *bf* because, by a direct analysis on the sample obtained by setting $bf = 1$, we saw that the maximum number of bridges adjacent to a node was 4. The values of the metrics obtained in this case are reported in Table 3.

From the analysis of this table we can observe that, as for the first four metrics, the obtained results are satisfying and comparable for each value of *bf*. This is a further confirmation that fixing $nf = 0.10$ allows BDS to cover all the social networks of the SIS in a satisfactory way. The only discriminant for *bf* seems to be *DB* because the increase in *bf* leads to an increase in the average node degree. This trend can be explained as follows. When a user has more adjacent bridges (less than 4, in our sample), for $bf = 0.25$, BDS selects only one of them. In this case, with the highest probability, the selected bridge will be the one with the lowest degree (see Line 13 in Algorithm 2). In the same case, if $bf = 1$, then BDS selects all adjacent bridges and, therefore, also those having the highest degrees. From the assortativity property [48], it is well known that high-degree users are often connected with other high-degree users. All these facts imply that the average degree of the sampled nodes increases. This explains the seemingly "strange" decrease in *BR* observed when *bf* increases. In fact, because the number of adjacent bridges is very limited, when the number of adjacent nodes increases, the fraction of bridges present in it (i.e., *BR*) decreases. All these reasonings suggest that an increase in *bf* causes an increase in *DB* and a decrease in *BR*. All the other metrics do not show significant variations. For this reason, in these experimental campaigns, we fixed *bf* to 0.25.

Table 4 Performances of BDS for different values of *btf*

$nf = 0.10$, $bf = 0.25$,	$btf = 0.00$	$btf = 0.25$	$btf = 0.50$	$btf = 0.75$	$btf = 1.00$
Bridge Ratio	0.0771	0.0965	0.0715	0.07612	0.0705
Crossing	245	286	226	258	183
Covering	4	4	4	4	4
Unbalancing	20.1134	13.5103	9.9273	11.8239	14.8959
Degree Bias	140.4817	19.0145	32.8467	48.7946	29.8698
Twitter Avg. Deg.	*40.2287*	*42.0383*	*39.7203*	*39.0245*	*39.6141*
YouTube Avg. Deg.	*11.8225*	*11.9278*	*11.1113*	*11.5115*	*11.3697*
Flickr Avg. Deg.	*379.377*	*129.5833*	*118.1566*	*124.1844*	*133.0608*
LiveJournal Avg. Deg.	*97.9286*	*68.6234*	*104.7473*	*136.0927*	*92.8981*

Impact of *btf*

In this experimental campaign, we fixed *nf* to 0.10 and *bf* to 0.25. We considered the following values for *btf*: 0, 0.25, 0.50, 0.75, and 1. *btf* can be seen as a filter on the bridge degrees. In particular, if $btf = 0$, then there is no constraint on the degrees of the bridges to select. If $btf = 1$, then BDS behaves as MH and, therefore, favors the selection of those bridges whose degree is lower than or equal to that of the current node. The other bridges are selected with a probability that decreases with the increase in their degree. The values of the metrics measured in this experimental campaign are reported in Table 4.

From the analysis of this table, it is evident that, when $btf = 0$, *DB* is high. This can be explained by the fact that, in this case, all bridges (even those with very high degree) may be equally selected. For the other values of *btf*, the overall performances of BDS do not present significant differences because all these values allow high-degree bridges to be filtered out. From a direct analysis on our samples we verified that the average degree of bridges is at most four times that of non-bridges. Setting $btf = 0.25$ allows that, even in the worst case (i.e., $p = 1$ in Line 13 of Algorithm 2), the bridges having a degree lower than or equal to the average degree of non-bridges are generally selected. This way, high-degree bridges are unfavored whereas the others are highly favored. For this reason, in these experimental campaigns, we fixed *btf* to 0.25.

3.2.4 Average Crawling Quality

So far we have analyzed the behavior of BDS w.r.t. the five metrics separately considered. To compare our strategy with the other three ones, it is more important to study their "overall" behavior by using the metric *ACQ*, which aggregates all the five metrics considered previously. Here, we have to deal with the problem of setting the coefficients of the linear combination, namely $w_{BR}, w_{CR}, w_{CV}, w_{UB}$, and w_{DB}, present in the definition of *ACQ*.

Table 5 *ACQ* for the different parameter configurations of BDS

Configuration	*ACQ* (same weights)	*ACQ* (different weights)
nf = 0.02 bf = 0.25 btf = 0.25	0.42	0.36
nf = 0.10 bf = 0.25 btf = 0.00	0.48	0.42
nf = 0.10 bf = 0.25 btf = 0.25	0.84	0.86
nf = 0.10 bf = 0.25 btf = 0.50	0.67	0.57
nf = 0.10 bf = 0.25 btf = 0.75	0.66	0.57
nf = 0.10 bf = 0.25 btf = 1.00	0.64	0.55
nf = 0.10 bf = 0.50 btf = 0.25	0.68	0.63
nf = 0.10 bf = 0.75 btf = 0.25	0.68	0.59
nf = 0.10 bf = 1.00 btf = 0.25	0.64	0.57
nf = 0.25 bf = 0.25 btf = 0.25	0.73	0.67
nf = 0.50 bf = 0.25 btf = 0.25	0.46	0.44

We start by assigning the same weight to all metrics, i.e., we set $w_{BR} = w_{CR} = w_{CV} = w_{UB} = w_{DB} = 0.2$. Then, we measure the value of *ACQ* for all the configurations of the parameters of BDS examined in Tables 2, 3 and 4. Obtained values are reported in the second column of Table 5. From the analysis of this column, we can verify that the configuration of BDS that guarantees the best tradeoff among the various metrics is $nf = 0.10$, $bf = 0.25$, $btf = 0.25$.

Observe that, at the beginning of Sect. 3.2 we showed that the defined metrics are not completely independent of each other. In fact, *BR* influences *CR* and *CV*, whereas *CR* influences *CV* and *UB*. As a consequence, it is reasonable to associate different weights with the various metrics by assigning the higher values to the most influential ones. To determine these values, we use an algorithm that takes inspiration from the Kahn's approach for topological sorting of graphs [29]. In particular, we first construct the *metric Dependency Graph*. It has a node n_{M_i} for each metric M_i. There is an edge from n_{M_i} to n_{M_j} if the metric M_i influences the metric M_j. A weight is associated with each node. Initially, we set all the weights to 0.20 (Fig. 1). We start from a node having no outgoing edges and split its weight (in equal parts) among itself and the nodes it depends on.[1] Then, we remove all the incoming edges. We repeat the previous tasks until all the nodes of the graph have been processed. By applying this approach we obtain the following configuration of weights: $w_{BR} = 0.45$, $w_{CR} = 0.18$, $w_{CV} = 0.07$, $w_{UB} = 0.10$, $w_{DB} = 0.20$. Observe that the node processing order is not unique because more than one node with no outgoing edge exists. However, it is easy to verify that the final metric weights returned by our algorithm do not depend on the adopted node processing order.

[1] Clearly, if a node has no incoming edges, it maintains its weight.

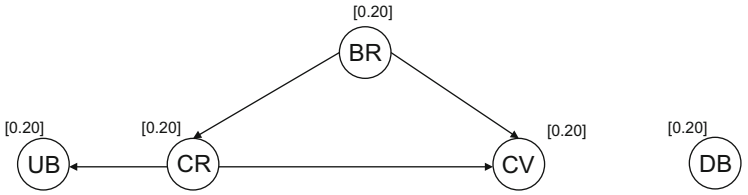

Fig. 1 The Dependency Graph concerning our metrics

Table 6 Values of *ACQ* for the different crawling techniques

Technique	*ACQ* (same weights)	*ACQ* (different weights)
MH	0.34	0.26
BFS	0.18	0.12
RW	0.08	0.03
BDS	0.84	0.86

Now, we measure *ACQ* with this new weight setting. The obtained results are reported in the third column of Table 5. Also in this experiment, the setting $nf = 0.10$, $bf = 0.25$, $btf = 0.25$ of the parameters of BDS shows the best performance.

However, with regard to this result, there are applications in which some metrics are more important than the others. BDS is highly flexible and, in these cases, allows the choice of the configuration that favors those metrics. For instance, if a user performs link mining in a SIS, the most important metric is *CR* because it is an index of the number of links between different social networks present in the crawled sample. By contrast, *DB* does not appear particularly relevant. In this case, the configuration $nf = 0.25$, $bf = 0.25$, $btf = 0.25$, is chosen because it guarantees the maximum *CR* even though the corresponding Bias Degree is quite high (see Table 2). As a second example, if it is desired a crawled sample in which all the social networks of the SIS are represented in the most uniform way, it is suitable to adopt the configuration $nf = 0.10$, $bf = 0.75$, $btf = 0.25$ that guarantees the best *UB* (see Table 3).

We now compare BDS, BFS, RW, and MH when they operate on a SIS. In this comparison, as for BFS, RW and MH we selected the overall values (see the last column of Table 1). As for BDS we adopted the configuration $nf = 0.10$, $bf = 0.25$, $btf = 0.25$. The results of this comparison, obtained by computing *ACQ* with the two weight settings, are reported in Table 6.

Interestingly enough, even the lowest value of *ACQ* obtained for BDS (obtained with the configuration $nf = 0.02, bf = 0.25, btf = 0.25$) is higher than that obtained for MH and much higher than those obtained for BFS and RW. This shows that BDS guarantees always the best performance.

From the analysis of these values and of those reported in Tables 1, 2, 3, and 4, it clearly emerges that, when operating on a SIS, BDS highly outperforms the other approaches. The only exception is MH for *DB* because, according to [27, 36], 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 *DB*, and the consideration that our crawling strategy, in absence of bridges, can be located between BFS and MH, producing intra-social-network results that reasonably cannot differ significantly from the above strategies.

4 Experiences

As pointed out in the introduction, the second main purpose of this paper is to exploit BDS for investigating the main features of bridges and SISs. This section is devoted to this analysis and is organized in such a way that each subsection investigates a specific aspect of SISs, namely the degree of bridges and non-bridges, the relationships between bridges and power users, the possible existence of a bridge backbone and the analysis of bridge centrality.

To perform the analyses of this section, we collected ten samples using BDS. We performed each investigation described below on each sample and, then, we averaged the obtained values on all of them. Therefore, each measure reported below is the average of the values obtained on each sample.

4.1 Distributions of Bridge and Non-bridge Degrees

In this section, we analyze the distributions of node degrees. For this purpose, we compute the Cumulative Distribution Function (CDF) of the degree of bridges and non-bridges. This function describes the probability that the degree of a node is less than or equal to a given value x. The CDFs for bridges and non-bridges are shown in Fig. 2.

By analyzing this figure, we can see that, fixed a degree d, the probability that a bridge has more than d contacts is higher than that of a non-bridge, for any d. As a consequence, we can state that a bridge has more contacts than a non-bridge, in average.

Again, observing the CDF trend for both bridges and non-bridges, it seems that the corresponding degrees follow a power law distribution. To verify this conjecture, in Fig. 3 we plot the Probability Distribution Function (PDF) of the degree of bridges and non-bridges. A visual analysis of the PDF trend already confirms our conjecture. To refine our analysis, we compute the best power law fit using the maximum likelihood method [16]. Table 7 shows the estimated power law coefficients, along with the Kolmogorov-Smirnov goodness-of-fit metrics, for the distributions into consideration. In particular, α is the exponent of the theoretical power law function

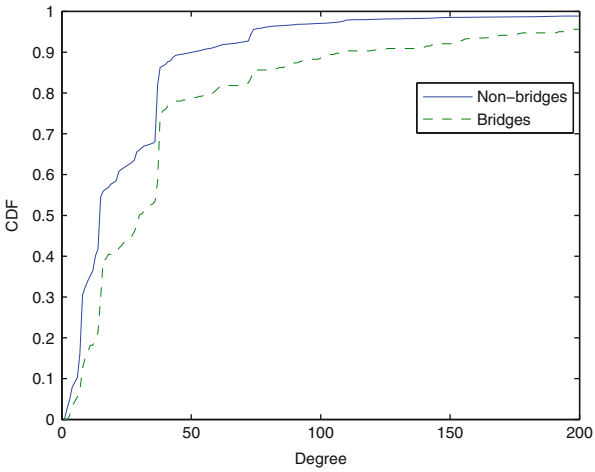

Fig. 2 Cumulative Distribution Function for bridges and non-bridges

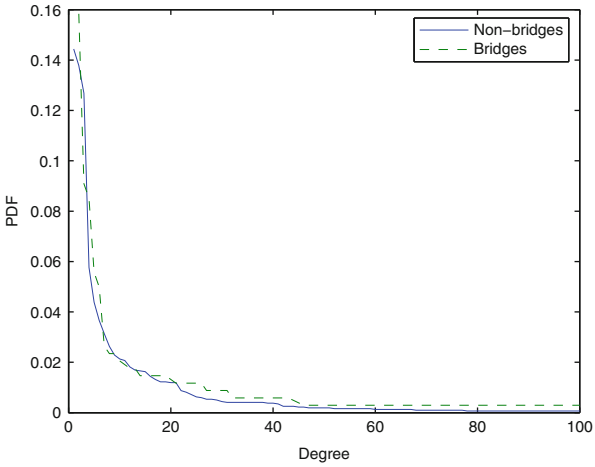

Fig. 3 Probability Distribution Function for bridges and non-bridges

Table 7 Power law coefficient estimation for the PDF of bridges and non-bridges

	α	D
Non-bridges	2.18	0.09
Bridges	2.31	0.15

that best approximates the real one, whereas D is the maximum distance between the theoretical function and the real one. The shown results, and in particular the low value of the Kolmogorov-Smirnov goodness-of-fit metric, confirm that the degrees of bridges and non-bridges follow a power law distribution.

To deepen our analysis, we compute the average degrees of bridges and non-bridges, along with the corresponding standard deviations, for each social network of the SIS. The obtained results are presented in Table 8. From the analysis of this table we can observe that: (1) the standard deviations of bridges and non-bridges are generally high; this can be explained by the power law distribution of PDF; (2) both

Table 8 Analysis of bridge and non-bridges degrees for the whole SIS and its social networks

	AVG		STD	
	Bridges	Non-bridges	Bridges	Non-bridges
YouTube	15.13	11.67	9.00	8.30
Flickr	315.47	97.01	834.41	119.96
LiveJournal	159.93	49.21	145.75	59.66
Twitter	44.21	41.75	20.84	19.73
All	66.69	26.82	282.10	41.13

the average degree and the degree standard deviation of bridges are higher than the corresponding ones of non-bridges for all the social networks of the SIS; in other words, this trend, valid for the SIS in the whole, is general and not specific for some social network.

In other words, BDS confirms the same results obtained in [11] with the other crawling techniques, and, in turn, this represents a further confirmation of Fact *(ii)* introduced in Sect. 3.

Continuing the analysis of Fig. 2, we can observe a relevant discontinuity of the CDF for both bridges and non-bridges around the degrees 35–40. Indeed, we have that the probability to find a bridge with less than 35 contacts is about 0.5, and that this probability becomes 0.75 for bridges with less than 40 contacts. The same trend occurs for non-bridges. This may be explained by considering that there exist two typologies of social network users. The former is composed by users who joined a social network for a short time, adding a limited (less than 40) number of friends. The latter refers to users who are active and, therefore, have an increasing number of contacts. The former typology raises the CDF values in the initial range (say 0–40), generating the observed discontinuity.

4.2 Bridges and Power Users

As seen in the previous section, bridges have an average degree higher than that of non-bridges. A question arises spontaneously: Are bridges power users? As a matter of fact, according to the definition given in [56], power users are nodes having a degree higher than the average degree of the other nodes. Indeed, if bridges were power users, for their detection it is possible to exploit the techniques for power user extraction already proposed in the literature. In the previous experiment, we measured that average degree of bridges is 66.69, whereas that of non-bridges is 26.82. The average degree of all nodes is 30.67, which is the reference value for classifying a node as power users. Looking at the average degrees, we may expect that bridges are actually power users. However, because degrees follow a power law distribution, this conjecture may be wrong.

To solve this question, we have to understand how much the set of power users and that of bridges overlap. Specifically, we denote by P the set of power users and

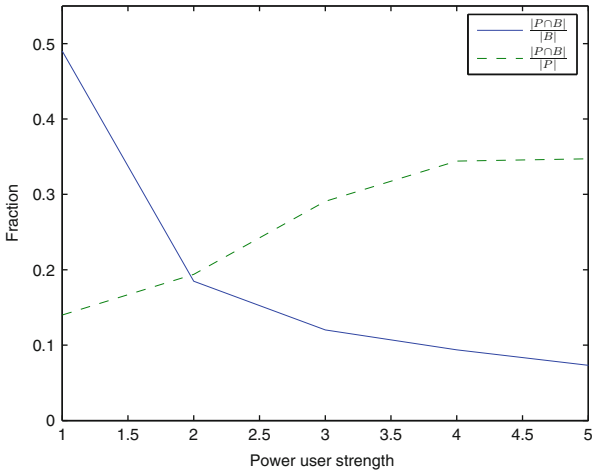

Fig. 4 Overlapping between bridge set and power user set

by B the set of bridges. Then, we measure the fraction of bridges that are power users and the fraction of power users that are bridges. The obtained results are: $\frac{|P\cap B|}{|B|} = 0.49$ and $\frac{|P\cap B|}{|P|} = 0.14$. They show that half of bridges are power users, whereas only few power users are bridges.

To better understand this phenomenon, we extend the concept of power user by introducing the notion of the strength of a power user. In particular, we say that a power user is an *s-strength power user* if its degree is s times higher than the average degree of nodes. Clearly, a standard power user corresponds to a 1-strength power user. Now, we compute $\frac{|P\cap B|}{|B|}$ and $\frac{|P\cap B|}{|P|}$ for increasing values of the strength of power user. The results of this experiment are shown in Fig. 4.

Here, it is possible to see that initially half of the bridges are power users. However, the percentage of bridges that are s-strength power users decreases as s increases. This allows us to conclude that bridges are not "strong" power users. Viceversa, the percentage of power users that are bridges increases as s increases. This allows us to conclude that the probability of finding a bridge among the strongest power users is higher than that of finding a bridge among the weak power users. However, this probability is never higher than 0.35.

In any case, both the (decreasing) trend of $\frac{|P\cap B|}{|B|}$ and the low values of $\frac{|P\cap B|}{|P|}$ allow us to conclude that there does not exist a meaningful correlation between bridges and power users.

4.3 Ties Among Bridges and Non-bridges

In this analysis, we aim at studying whether bridges have preferential ties among them, i.e., whether they are more likely to be connected to each other than to non-bridges. A possible way to carry out this verification is to compute the distribution of

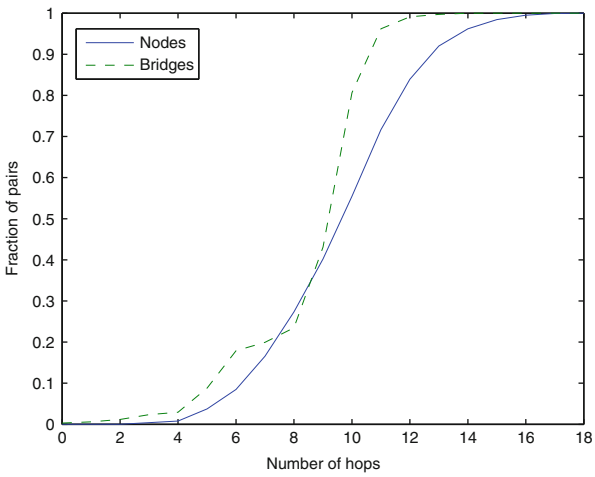

Fig. 5 Distribution of the lengths of the shortest paths among bridges and among nodes

the lengths of the shortest paths among bridges and among nodes in the SIS. Indeed, if these distributions are similar and the maximum lengths of the shortest paths connecting two nodes and two bridges are comparable, it is possible to conclude that no preferential tie exists among bridges. By contrast, if the maximum length of the shortest paths connecting two bridges is less than that of the shortest paths connecting two nodes and/or the distributions are dissimilar in the sense that the one of bridges raises much faster than the one of non-bridges, it is possible to conclude that bridges are likely to be connected to each other. The results of this experiment are shown in Fig. 5.

From the analysis of this figure, we can see that the distribution of bridge distance follows the same trend of that of node distance. Moreover, the effective diameter (90-th percentile of the distribution of the lengths of the shortest paths) measured for nodes and bridges is about 12 and 11, respectively. This allows us to conclude that no preferential connection favoring the link among bridges exists.

Interestingly enough, this experiment supplies us a further hint about the node to be used as seed in a crawling task for a SIS: Indeed, we cannot rely on a particular seed to enhance the percentage of bridges in a crawled sample, because a backbone among bridges does not exist.

4.4 Bridge Centrality

This experiment is devoted to analyze the centrality of bridges in a SIS. Centrality is one of the most important measures adopted in Social Network Analysis to investigate the features of nodes in a social network. Basically, there are four main centrality metrics, namely *degree*, *betweennes*, *closeness* and *eigenvector* [20]. In this experiment, we focus on *betweenness* because it computes the centrality of a

Table 9 Centrality of bridges and non-bridges

	Bridges	Non-bridges
Twitter	2,404	2,758
Flickr	643	652
LiveJournal	33	36
YouTube	3,779	5,685
All	23,100	12,156

node by quantifying how much it is important in guaranteeing the communication among other nodes and, therefore, how much it acts as bridge along the shortest paths of other nodes.

We expect that bridges have a high *betweenness* value in the whole SIS. In this experiment, we aim at verifying this intuition and, in the affirmative case, at studying if they maintain this property in the single social networks they joined.

For this purpose, we compute the *betweennes* of each node in our samples by means of *SNAP* (Stanford Network Analysis Platform) [56] and we average the corresponding values for bridges and non-bridges. The results are reported in Table 9.

By analyzing this table, we can observe that the intuition about the high *betweennes* of bridges in the whole SIS is fully confirmed. By contrast, in the single social networks, the values of *betweennes* of bridges are comparable or less than those of non-bridges. At a first glance, this result is unexpected because it appears immediate to think that a bridge can maintain its role of connector also in the single social networks joined by it. Actually, a more refined reasoning leads us to conclude that often bridges, just for their role, are at the borders of their social networks, and this partially undermines their capability to be central.

5 Conclusion

In this paper, first we have investigated the problem of crawling Social Internetworking Scenarios. We have started from the consideration that existing crawling strategies are not suitable for this purpose. In particular, we have analyzed the state-of-the-art techniques, which are BFS, RW and MH, showing experimentally that the above claim is true. On the basis of this result, by analyzing the reasons of the drawbacks of existing crawling strategies, we have designed a new one, called BDS (Bridge-Driven Search), specifically conceived for a SIS. We have conducted several experiments showing that, when operating in a SIS, BDS highly outperforms BFS, RW and MH, and arguing that BDS presents a good behavior also in intra-social-network crawling. Besides the overall conclusion mentioned above, we have seen that BDS is highly flexible as it allows a metric to be privileged over another one. After having validated BDS, we have exploited it to explore the emergent scenario of Social Internetworking from the perspective of Social Network Analysis. Being aware that the complete investigation of all the aspects of SISs is an extremely large task, we have

identified the most basic structural peculiarity of these systems, i.e. bridges, and we have deeply studied it. We argue that most of the knowledge about the structural properties of SISs, and possibly about the behavioral aspects of users, starts from the adequate knowledge of bridges, which are the structural pillars of SISs.

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 is the improvement of the BDS crawling strategy in such a way that the values of *nf*, *bf* and *btf* dynamically change during the crawling activity to adapt themselves to the specificities of the crawled SIS. Moreover, we plan to investigate the possible connections of our approach with the information integration ones, as well as to deal with the privacy issue arising when crawling SISs. Another important future development regards the exploitation of BDS (or its evolutions) to perform a deeper investigation of SISs. In this context, it appears extremely promising to apply Data Warehousing, OLAP and Data Mining techniques on SIS samples derived by applying BDS to derive knowledge patterns about SISs.

Acknowledgements This work has been partially supported by the TENACE PRIN Project (n. 20103P34XC) funded by the Italian Ministry of Education, University and Research.

References

1. Agarwal N, Galan M, Liu H, Subramanya S (2010) WisColl: collective wisdom based blog clustering. Inf Sci 180(1):39–61
2. Ahn YY, Han S, Kwak H, Moon S, Jeong H (2007) Analysis of topological characteristics of huge online social networking services. In: Proceedings of the international conference on world wide web (WWW'07), Banff, Alberta. ACM, New York, pp 835–844
3. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD'06), Philadelphia. ACM, New York, pp 44–54
4. Berlingerio M, Coscia M, Giannotti F, Monreale A, Pedreschi D (2010) Towards discovery of eras in social networks. In: Proceedings of the workshops of the international conference on data engineering (ICDE 2010), Long Beach. IEEE, Los Alamitos, CA, USA, pp 278–281
5. Berlingerio M, Coscia M, Giannotti F, Monreale A, Pedreschi D (2011) Foundations of multidimensional network analysis. In: Proceedings of the international conference on advances in social networks analysis and mining (ASONAM 2011), Kaohsiung. IEEE, Los Alamitos, CA, USA, pp 485–489
6. Berlingerio M, Coscia M, Giannotti F, Monreale A, Pedreschi D (2011) The pursuit of hubbiness: analysis of hubs in large multidimensional networks. J Comput Sci 2(3):223–237
7. Bonneau J, Anderson J, Danezis G (2009) Prying data out of a social network. In: Proceedings of the international conference on advances in social network analysis and mining (ASONAM'09), Athens. IEEE, Los Alamitos, CA, USA, pp 249–254
8. Brickley D, Miller L (2012) The friend of a friend (FOAF) project. http://www.foaf-project.org/
9. Buccafurri F, Lax G, Nocera A, Ursino D (2012) Crawling social internetworking systems. In: Proceedings of the international conference on advances in social analysis and mining (ASONAM 2012), Istanbul. IEEE Computer Society, Los Alamitos, pp 505–509

10. Buccafurri F, Lax G, Nocera A, Ursino D (2012) Discovering links among social networks. In: Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD 2012), Bristol. Lecture notes in computer science. Springer, Berlin, pp 467–482
11. Buccafurri F, Foti VD, Lax G, Nocera A, Ursino D (2013) Bridge analysis in a social internetworking scenario. Inf Sci 224:1–18
12. Carrington P, Scott J, Wasserman S (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
13. Catanese SA, De Meo P, Ferrara E, Fiumara G, Provetti A (2011) Crawling Facebook for social network analysis purposes. In: Proceedings of the international conference series on web intelligence, mining and semantics (WIMS'11), Sogndal. ACM, New York, pp 52–59
14. Chau DH, Pandit S, Wang S, Faloutsos C (2007) Parallel crawling for online social networks. In: Proceedings of the international conference on world wide web (WWW'07), Banff, Alberta. ACM, New York, pp 1283–1284
15. Cheng X, Dale C, Liu J (2008) Statistics and social network of Youtube videos. In: Proceedings of the international workshop on quality of service (IWQoS 2008), Enschede. IEEE, Los Alamitos, CA, USA, pp 229–238
16. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703
17. Dai BT, Chua FCT, Lim EP (2012) Structural analysis in multi-relational social networks. In: Proceedings of the international SIAM conference on data mining (SDM 2012), Anaheim. Omnipress, Madison, pp 451–462
18. De Meo P, Ferrara E, Fiumara G, Provetti A (2011) Generalized Louvain method for community detection in large networks. In: Proceedings of the international conference on intelligent systems design and applications (ISDA 2011), Cordoba. IEEE, Los Alamitos, CA, USA, pp 88–93
19. de Sola Pool I, Kochen M (1978) Contacts and influence. Soc Netw 1:5–51
20. Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1(3): 215–239
21. FriendFeed (2012). http://friendfeed.com/
22. Gathera (2012). http://www.gathera.com/
23. Ghosh R, Lerman K (2010) Predicting influential users in online social networks. In: Proceedings of the KDD international workshop on social network analysis (SNA-KDD'10), San Diego. ACM, New York
24. Google Open Social (2012). http://code.google.com/intl/it-IT/apis/opensocial/
25. Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the international conference on human factors in computing systems (CHI'09), Boston. ACM, New York, pp 211–220
26. Gilbert AC, Levchenko K (2004) Compressing network graphs. In: Proceedings of the international workshop on link analysis and group detection (LinkKDD'04), Seattle. ACM, New York
27. Gjoka M, Kurant M, Butts CT, Markopoulou A (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the international conference on computer communications (INFOCOM'10), San Diego. IEEE, Los Alamitos, CA, USA, pp 1–9
28. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
29. Kahn AB (1962) Topological sorting of large networks. Commun ACM 5(11):558–562
30. Kazienko P, Musial K, Kukla E, Kajdanowicz T, Bródka P (2011) Multidimensional social network: model and analysis. In: Proceedings of the international conference on computational collective intelligence (ICCCI 2011), Gdynia. Springer, Berlin, pp 378–387
31. Kleinberg J (2008) The convergence of social and technological networks. Commun ACM 51(11):66–72
32. Korolova A, Motwani R, Nabar SU, Xu Y (2008) Link privacy in social networks. In: Proceedings of the ACM international conference on information and knowledge management (CIKM'08), Napa Valley. ACM, New York, pp 289–298

33. Krishnamurthy V, Faloutsos M, Chrobak M, Lao L, Cui JH, Percus A (2005) Reducing large internet topologies for faster simulations. In: Proceedings of the international conference on networking (Networking 2005), Waterloo, Ontario. Springer, Berlin, pp 165–172
34. Krishnamurthy B, Gill P, Arlitt M (2008) A few chirps about Twitter. In: Proceedings of the first workshop on online social networks, Seattle, pp 19–24
35. Kumar R, Novak J, Tomkins A (2010) Structure and evolution of online social networks. In: Link mining: models, algorithms, and applications, Springer, New York, pp 337–357
36. Kurant M, Markopoulou A, Thiran P (2010) On the bias of BFS (Breadth First Search). In: Proceedings of the international teletraffic congress (ITC 22), Amsterdam. IEEE, Los Alamitos, CA, USA, pp 1–8
37. Lee SH, Kim PJ, Jeong H (2006) Statistical properties of sampled networks. Phys Rev E 73(1):016102
38. Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD'06), Philadelphia. ACM, New York, pp 631–636
39. Li YM, Lai CY, Chen CW (2011) Discovering influencers for marketing in the blogosphere. Inf Sci 181(23):5143–5157
40. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102(33):11623–11628
41. Lovász L (1993) Random walks on graphs: a survey. In: Combinatorics, Paul Erdos is eighty, vol 2, no 1, Springer, Heidelberg, Germany, pp 1–46
42. Mathioudakis M, Koudas N (2009) Efficient identification of starters and followers in social media. In: Proceedings of the international conference on extending database technology: advances in database technology (EDBT '09), Saint Petersburg. ACM, New York, pp 708–719
43. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the ACM SIGCOMM international conference on internet measurement (IMC'07), San Diego. ACM, New York, pp 29–42
44. Mislove A, Koppula HS, Gummadi KP, Druschel F, Bhattacharjee B (2008) Growth of the Flickr social network. In: Proceedings of the international workshop on online social networks (WOSN'08), Seattle. ACM, New York, pp 25–30
45. Monclar R, Tecla A, Oliveira J, de Souza JM (2009) MEK: using spatial–temporal information to improve social networks and knowledge dissemination. Inf Sci 179(15):2524–2537
46. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328(5980):876–878
47. Musiał K, Juszczyszyn K (2009) Properties of bridge nodes in social networks. In: Proceedings of the international conference on computational collective intelligence (ICCCI 2009), Wroclaw. Springer, Berlin, pp 357–364
48. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701
49. Onnela JP, Reed-Tsochas F (2010) Spontaneous emergence of social influence in online systems. Proc Natl Acad Sci 107(43):18375
50. Perer A, Shneiderman B (2006) Balancing systematic and flexible exploration of social networks. IEEE Trans Vis Comput Graph 12(5):693–700
51. Power.com (2012). http://techcrunch.com/2008/11/30/powercom-for-social-networking-power-users/
52. Rafiei D, Curial S (2005) Effectively visualizing large networks through sampling. In: Proceedings of the IEEE visualization conference 2005 (VIS'05), Minneapolis. IEEE, Los Alamitos, CA, USA, p 48
53. Rasti AH, Torkjazi M, Rejaie R, Stutzbach D (2008) Evaluating sampling techniques for large dynamic graphs. Univ. Oregon, Tech. Rep. CIS-TR-08-01
54. Romero DM, Galuba W, Asur S, Huberman BA (2011) Influence and passivity in social media. In: Proceedings of the international conference on world wide web (WWW'11), Hyderabad. ACM, New York, pp 113–114
55. Song X, Chi Y, Hino K, Tseng B (2007) Identifying opinion leaders in the blogosphere. In: Proceedings of the ACM international conference on information and knowledge management (CIKM'07), Lisbon. ACM, New York, pp 971–974

56. Stanford Network Analysis Package (2012). http://snap.stanford.edu/snap/
57. Stutzback D, Rejaie R, Duffield N, Sen S, Willinger W (2006) On unbiased sampling for unstructured peer-to-peer networks. In: Proceedings of the international conference on internet measurements, Rio De Janeiro. ACM, New York, pp 27–40
58. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32(4):425–443
59. Wilson C, Boe B, Sala A, Puttaswamy KPN, Zhao BY (2009) User interactions in social networks and their implications. In: Proceedings of the ACM European conference on computer systems (EuroSys'09), Nuremberg. ACM, New York, pp 205–218
60. Wu A, DiMicco JM, Millen DR (2010) Detecting professional versus personal closeness using an enterprise social network site. In: Proceedings of the international conference on human factors in computing systems (CHI'10), Atlanta. ACM, New York, pp 1955–1964
61. XFN - XHTML Friends Network (2012). http://gmpg.org/xfn
62. Ye S, Lang J, Wu F (2010) Crawling online social graphs. In: Proceedings of the international Asia-Pacific web conference (APWeb'10), Busan. IEEE, Los Alamitos, CA, USA, pp 236–242