Information Sciences 224 (2013) 1–18

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Bridge analysis in a Social Internetworking Scenario

Francesco Buccafurri [a,*], Vincenzo Daniele Foti [b], Gianluca Lax [a], Antonino Nocera [a], Domenico Ursino [a]

[a] DIMET Dept., University of Reggio Calabria, Feo di Vito, 89122 Reggio Cal., Italy
[b] SDET, Microsoft 3850 148th Ave. NE, 98052 Redmond, WA, USA

ARTICLE INFO

Article history:
Received 29 November 2011
Received in revised form 5 October 2012
Accepted 10 October 2012
Available online 26 October 2012

Keywords:
Social Network
Social Network Analysis
Social Internetworking Scenario
Bridge users
Crawling strategies

ABSTRACT

The rapid development of the number and the size of Online Social Networks (OSNs) makes the analysis of Social Internetworking Scenarios (SISs) extremely challenging. In a SIS, a user can join multiple OSNs and two users can interact with each other even though they joined different OSNs and did not know each other. While OSNs have been extensively studied in the last years, the most peculiar aspects of Social Internetworking Scenarios have not been yet investigated, especially from the Social Network Analysis perspective. Our paper tries to give a first important contribution in this field by deeply studying the core elements of a SIS, i.e., *bridges*. Bridges are those users who joined more OSNs and allow users of different OSNs to cooperate. We investigate the main features of this category of users by means of a Social Network Analysis campaign. In particular, we define several specific crawling strategies and extract several samples from a SIS by applying each of them. The experimental results define a clear "identikit" of bridges allowing us to state a number of non-trivial conclusions about their role in a SIS.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Online Social Networks (OSNs, for short) represent one of the main actors of the Web 2.0 and have been showing an enormous growth in the last years. They attracted the interest of many researchers from disparate fields, such as computer science, psychology and sociology [39]. Many researchers started to collect large amounts of data from OSNs and to apply techniques of classical Social Network Analysis [20] on them. The results they obtained are numerous and extremely interesting (see, for instance, [24,48,28,57]). Other researchers modeled an OSN as a graph and investigated its structural properties also to infer knowledge about user behavior [31,17]. They are based on the intuition that there is a strong correspondence between the user behavior in an OSN and the structural properties of the corresponding graph. This intuition is confirmed by many past studies, as shown in Section 2.

An important aspect to consider is that nowadays users tend to spread their activities among more OSNs and, often, to show a different behavior in different OSNs [36]. As a consequence, different Social Networks are interconnected with each other, thus resulting in a global graph whose structural features could be in principle very different from the ones of each Social Network seen as a graph. This complex topology represents the substrate of an emergent scenario, called *Social Internetworking Scenario* (*SIS*), where a user can join multiple OSNs and two users can interact with each other even though they joined different OSNs and did not know each other. This concept is very recent and only a few *commercial* attempts to

* Corresponding author.
E-mail addresses: bucca@unirc.it (F. Buccafurri), vfoti@microsoft.com (V.D. Foti), lax@unirc.it (G. Lax), a.nocera@unirc.it (A. Nocera), ursino@unirc.it (D. Ursino).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.10.021

implement Social Internetworking Systems have been proposed (think, for instance, of Google Open Social [4], Power.com [6], Gathera [3] and Friendfeed [2]).

Despite the great attention given by scientists towards Social Networks, Social Internetworking Scenarios have been little investigated in the scientific literature [27], also due to their young age. Some papers focus on *cross-folksonomies* [37,60]. They analyze the tagging behavior of users in multiple folksonomies and try to relate information about these behaviors. Other ones, such as [9,59], assume users can join *heterogenous social systems* (e.g., a folksonomy and a blogging platform in [59], or Social Networks like Facebook, and social media, like Flickr, in [9]). Their goal is to *aggregate* user information in such a way as to build a *global* user profile. However, no relevant work studying the most peculiar aspects of SIS's from the Social Network Analysis perspective exists in the scientific literature. In other words, the comprehension of even basilar structural properties of SIS's is still a fully open problem.

Our goal is to provide a first important contribution in this setting by deeply studying the core elements of a SIS, i.e. *bridges*. Bridges are those users who joined more OSNs and allow users of different OSNs to cooperate. A bridge generally operates at the borders of two or more OSNs, is very esteemed in them, is very active in her interactions with other users and, therefore, can favor information exchange among different OSNs. As a consequence, bridges certainly represent the most basic structural aspect of a SIS, which requires a specific careful analysis.

Our paper attacks this problem by following an experimental approach based on suitable crawling tasks on a real-life SIS. In particular, we run the following steps: *(i)* extraction of data concerning the connections among user accounts in a set of OSNs and *(ii)* identification and analysis of bridges.

In order to carry out the first step it is necessary to derive the connections explicitly declared by users. Furthermore, it is necessary to consider not only connections among accounts of the same OSN but also connections among accounts of different OSNs. For this task, technological standards encoding human relationships, such as XFN (XHTML Friends Network) [8] and FOAF (Friend-Of-A-Friend) [7], can be exploited for those cases when users explicitly declared these relationships. In all the other cases, human relationships should be inferred by suitable approaches (for instance, the one described in [19]). Our extraction task considers these standards but, instead of handling and processing them directly, it exploits the functionalities provided by *Google Social Graph (GSG)*, a tool recently proposed by Google [5]. It is worth noting that the above extraction step is not a simple task, since the crawling strategies adopted in the classical context cannot be trivially run in a SIS context by assuming that they capture its specific peculiarities. Therefore, both an extension of classical crawling strategies to SIS's and/or the design of new ad hoc strategies are necessary. From this point of view, our paper offers a first interesting contribution since, besides the results of the analysis on bridges, it includes a first investigation about the issue of crawling SIS's.

But the main contribution of our paper concerns step *(ii)* above. The analysis of bridges (whose identification, among the data extracted in step *(i)*, is straightforward), 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, by running a large number of experiments over large data sets driven by the following questions:

1. Is there a sort of "backbone" among bridges, aiming at favoring the direct links among them?
2. Which probabilistic distribution is followed by the bridge degree?
3. Which is the relationship among the fraction of bridges, the average node degree, and its standard deviation in the OSNs of a SIS?
4. What about the average degree of bridges? Is there some difference with non-bridges?
5. Is there a correlation between bridges and power users?
6. What about the number of OSNs typically connected by bridges?

The result of our analysis provides an answer to each question above with a strong experimental support and discovers even unexpected conclusions about bridges and, in general, a complete knowledge of this crucial elements of Social Internetworking Scenarios.

Summarizing, the main contributions of this paper are the following:

- we extended two classical crawling techniques (namely, BFS and MH) in order to make them more suitable to operate in a SIS;
- we tested five crawling techniques (including the two old ones) and derived several data sets about SIS's that we made available to the Social Network Analysis community for further experiments; and
- we defined the main characteristics of bridges, which represent the key factor in a SIS, by giving an answer to the above questions.

The plan of this paper is as follows: Section 2 describes related literature. Section 3 presents the crawling strategies we have implemented and used, the test bed we have adopted and the basic characteristics of collected data. Bridge analysis, which represents the core of this paper, is presented in Section 4. Finally, in Section 5 we draw our conclusions and identify some future directions of our research.

Throughout the paper, with a little abuse of notation, we indifferently talk of users referring to both their human role and the corresponding nodes of a graph.

2. Related literature

Initially, studies on Social Networks attracted mainly sociologists. For instance, [61] introduced the 6-degrees of separation and the small-world theories. The effects of these theories are analyzed in [29]. Ref. [35] 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 also attracted computer scientists and, as pointed out in the introduction, 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 [39]. However, to the best of our knowledge, no study that analyzes the main features of bridges in a SIS was presented in the past.

In [44], the authors discuss the problem of sampling from large graphs. They aim at answering questions like: *(i)* which sampling method to use; *(ii)* how small can the sample size be; *(iii)* how to scale up the measurements of the sample to get estimates for the larger graph; and *(iv)* how success can be measured. For this purpose, they consider several sampling methods and check the goodness of their sampling strategies on several different datasets. Other approaches that consider the same problem can be found in [32,55,41,12]. Specifically, [32,55] use sampling to improve the visualization of a graph. Ref. [41] develops methods to produce a small realistic sample from a large real network. The authors of this work prove that some of these methods maintain key properties of the initial graph also with a sample size down to 30%. Ref. [12] considers different graph generation algorithms and, for each of them, analyzes separability and stability properties.

Ref. [64] investigates the OSN graph crawling problem. Its analysis aims at answering questions like: *(i)* how fast crawlers into consideration discover nodes/links; *(ii)* how different OSNs and the number of protected users affect crawlers; *(iii)* how major graph properties are studied. All these investigations are performed by analyzing samples derived from four OSNs, namely Flickr, LiveJournal, Orkut and YouTube. Another approach that considers the same problem can be found in [18].

In [49], 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 ones contain a densely connected core of high-degree nodes. In [22], the authors describe a parallel crawler based on Breadth First Search (BFS) and operating on eBay. Studies about how an attacker discovers a social graph can be found in [40,16]. The sole purpose of the attacker is to maximize the number of nodes/links it can discover. As a consequence, these two papers do not examine other issues such as biases. In [43], the authors analyze 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. In particular, they quantify the bias of BFS in estimating the node degree w.r.t. the fraction of sampled nodes. Furthermore, they show that this bias can be corrected reasonably well even in case of very small sample size.

In [42], 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 [13], the authors investigate the main features of groups in LiveJournal and propose models that represent the growth of user groups over time. The authors of [11] analyze the topological properties of Cyworld (a South Korean OSN), Myspace and Orkut. They obtained data of the whole Cyworld and some samples of the other two OSNs. For the crawling task they use BFS. They show that this technique is capable of accurately approximating some network properties (e.g., degree distribution and correlation) even with small samples. In [46], 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. Ref. [20] 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, [26] 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 [54], 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 [34], the authors present an analysis of the Facebook friendship graph devoted to estimate any user property and some topological properties. In this activity they examine and compare several candidate crawling strategies, namely BFS, Random Walk, Metropolis-Hasting Random Walk and Re-Weighted Random Walk. The authors investigate also diagnostics to assess sample quality during the data collection process. In [21], 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 [62]. 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.

Ref. [50,30,53,56] present some approaches for the identification of influential users, i.e. users capable of stimulating other ones to join OSN activities and/or to actively operate in them. In [58,10,45], the authors suitably model the blogosphere to perform leader identification. In [47], 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 [52] 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 the one adopted in our paper. The authors base their analysis on the study of the theoretical properties of their model. In [33], 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 [63] 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 a user on an OSN reflects the strength of her relationships with other users w.r.t. several factors, like profile commenting and mutual connections.

Finally, [14,15,25,51,38] present some approaches in the field of multidimensional networks. These networks can be seen as a specific case of a Social Internetworking Scenario in which each social network is specific for one kind of relationship and social networks strongly overlap. Multidimensional social networks are also known as multislice networks in the literature [51]. The first definition of a model for multidimensional networks has been proposed in [14], where a set of measures to extract useful knowledge on multidimensional networks, along with their experimental validation on real-life datasets, is presented. Another model for multidimensional social networks has been proposed in [38]. This model, which resembles the concept of Data Warehouse, defines three main dimensions: relation layers, time windows and groups. A view describes the state of one social group, linked by only one kind of relationship (one layer), derived from within only one time period. The paper discusses also several aggregation possibilities and illustrates some possible uses of the proposed model. The concept of hub, intended as a node highly connected to the other ones, in multidimensional social networks is studied in [15]. The authors redefine the concept of degree in the multidimensional context and introduce a new class of measures, called Dimension Relevance, aiming at analyzing the importance of different dimensions for the capability of a node to act as a hub. Experiments performed on real networks showed that such hubs really exist and they can be found and studied by using suitable measures of interplay of the different dimensions. In the same paper, the authors show how to detect the most influential dimensions that cause the different hub behaviors. In [25], a Generalized Stochastic Block Model (GSBM, for short) for performing positional and role analysis on multi-relational networks is proposed. The authors adopt several Multivariate Probability Distribution Functions to model different kinds of multi-relational network. They experimentally show that their model is able to discover the ground truth grouping in synthetic networks and to predict relationships in real networks.

3. Methods and data for analysis

In this section, we first describe the crawling techniques exploited to collect real-life datasets from the considered SIS, and then we report the basic statistics of collected data.

3.1. Adopted crawlers

As pointed out in the introduction, in order to perform our analyses on bridges, we have to extract some samples from a SIS. These samples registered connections among user accounts. For carrying out the sampling activity, we have to realize some crawlers capable of operating on a SIS, instead of on a single OSN. Specifically, these crawlers have to be able to extract not only connections among the accounts of different users in the same OSN but also connections among the accounts of the same users in different OSNs.

Two standards encoding human relationships are generally exploited to define these last connections. The former is XFN (XHTML Friends Network) [8]. XFN simply uses an attribute, called `rel`, to specify the kind of relationship between two users. Some possible values of `rel` are `friend`, `contact`, `co-worker`, `parent`, and so on. A (presumably) more complex alternative to XFN is FOAF (Friend-Of-A-Friend). A FOAF profile is essentially an XML file describing a person, her links with other people and the links to the objects created by her. It is worth pointing out that the technicalities concerning these two standards are not to be handled manually by the user. As a matter of fact, each OSN has suitable mechanisms to automatically manage them in a way transparent to the user, who has just to specify her relationships in a user-friendly fashion. In order to simplify the management of these standards, our crawlers adopt the functionalities provided by Google Social Graph (hereafter, GSG) [5] which is capable of directly handling them in an efficient way. GSG is a tool proposed by Google to find information among people of those OSNs handling XFN and FOAF protocols. Given a user u who joins several OSNs, GSG returns both: *(i)* a list of public URLs associated with her and *(ii)* a list of publicly declared connections between her and other users. GSG is provided with a suitable method called `lookup`, allowing us to explore connections among users. Given a user u represented by a node n_u, a call to `lookup` on n_u generates two lists of nodes. The former contains the nodes pointing to n_u, whereas the latter consists of the nodes referenced by n_u. The answer to a `lookup` call is a *JSON* file. JSON is a lightweight data exchange format which can be easily understood by humans and processed by machines. Finally, a nice feature of GSG is its capability of handling `me` edges, i.e., of identifying accounts (often located in different Social Networks) which refer to the same individual. Also the management of the technicalities concerning `me` edges is automatically performed by the involved OSNs. The user must at most specify the corresponding URLs in a friendly fashion.

As stated in the introduction, sometimes `me` connections are not explicitly specified by users. However, in the literature, several approaches for detecting implicit `me` connections have been proposed. For instance, the approach described in [19] is

based on a notion of node similarity obtained by combining two contributions: a string similarity between the associated user names, and a contribution based on a suitable recursive notion of common-neighbor similarity. The latter is extremely important because it allows synonymy and homonymy errors to be detected and avoided.

As pointed out in Section 2, several crawling strategies for a single OSN have been proposed in the literature. For each of them their features have been investigated. This investigation showed that there exists no crawling strategy which is always better than the other ones. By contrast, each technique could be the optimal one for a specific set of analyses.

In order to carry out our analyses we needed crawling strategies capable of operating on a SIS, instead of on a single OSN. Clearly, it was not possible to a priori rely on the fact that the behavior of a crawling strategy in a single OSN does not change when it operates in a SIS. A verification of this issue was in order. On the basis of this reasoning, it appeared convenient to exploit several crawling strategies; some of them are classical ones extended to a SIS, whereas other ones have been defined to take into account the typical aspects of a SIS. As a consequence, our paper provides a further contribution, in addition to the main one consisting in bridge analysis. Indeed, it starts to investigate the features and the peculiarities of several crawling strategies in a SIS (which, as previously pointed out, is quite a specific scenario).

The crawling strategies we have implemented are the following:

- *BFS*: It implements the classical Breadth First Search visit (stopped after a suitable number n_{it} of iterations).
- BFS^B: Its underlying philosophy is derived from *BFS* by including some features which favor the presence of bridges in the sample. In particular, it starts with a *BFS* visit on a seed node *s*. When it finds a bridge, it inserts this bridge in a set called *B*. During each iteration, if *B* is empty, it continues the current *BFS* visit. By contrast, if *B* is not empty, it continues the current *BFS* visit with a probability $p = 0.85$, whereas, with a probability $1 - p$, it starts a new *BFS* visit taking a bridge of *B* as seed. It stops after n_{it} iterations. The pseudocode of this algorithm is shown in Algorithm 1.

Algorithm 1. BFS^B

Input: *s*: a seed node
Output: *SeenNode*, *VisitedNode*: a set of nodes
Constant n_{it} {The number of iterations}
Variable *v*: a node
Variable *p*: a number in the real interval (0,1)
Variable *B*: set of nodes
Variable *NodeList*: an ordered list of nodes accessed according to a FIFO policy

```
1: SeenNode := ∅, VisitedNode := ∅, B := ∅, NodeList := ∅
2: insert all the nodes adjacent to s into NodeList
3: insert s into SeenNode and VisitedNode
4: insert all the nodes adjacent to s into SeenNode
5: for i := 1 to n_it do
6:    generate uniformly at random a number p in the real interval (0,1)
7:    if (p < 0.15) and (B ≠ ∅) then
8:       extract uniformly at random a node v from B
9:       NodeList := ∅
10:   else
11:      extract a node v from NodeList
12:      if v is a bridge then
13:         insert v into B
14:      end if
15:   end if
16: insert all the nodes adjacent to v into NodeList
17: insert v into VisitedNode
18: insert all the nodes adjacent to v into SeenNode
19: end for
```

- *M*: It implements the Metropolis-Hasting Random Walk that proved to perform very well in past analyses of single OSNs [34]. This crawling strategy has been conceived to unfavor power users who, instead, are favored by BFS-like crawling strategies that, owing to this fact, can present bias in some network parameters (e.g., the average number of contacts per user) [43].[1]

[1] It is worth pointing out that BFS proved to perform very well in estimating other parameters, like clustering coefficient [43].

M starts its visit from a seed node s. During each iteration it considers the currently visited node v and randomly selects a node w from the neighbors of v. Then, it randomly generates a number p belonging to the real interval $(0,1)$. If $p \leqslant \frac{k_v}{k_w}$, where k_v(k_w, resp.) is the outdegree of v (w, resp.), then it moves from v to w. Otherwise, it stays in v. It terminates after n_{it} iterations. The pseudocode of this algorithm is shown in Algorithm 2. Observe that the higher the degree of a node, the higher the probability that M discards it.

Algorithm 2. M

Input s: a seed node
Output *SeenNode*, *VisitedNode*: a set of nodes
Constants n_{it} {The number of iterations}
Variable v, w: a node
Variable p: a number in the real interval (0,1)
1: $SeenNode := \emptyset$, $VisitedNode := \emptyset$
2: insert s into *SeenNode* and *VisitedNode*
3: insert all the nodes adjacent to s into *SeenNode*
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: let k_v and k_w be the outdegree of v and w, respectively
9: **if** $(p \leqslant \frac{k_v}{k_w})$ **then**
10: $v = w$
11: insert w into *VisitedNode*
12: insert all the nodes adjacent to w into *SeenNode*
13: **end if**
14: **end for**

- M^B: Its underlying philosophy is similar to the one of M. However, it includes some features which favor the presence of bridges in the sample. It starts its visit from a seed node s. During each iteration it considers the currently visited node v and randomly selects a node w from the neighbors of v. If w is a bridge, then it moves from v to w; otherwise, it behaves as M. The pseudocode of this algorithm is identical to the one of M, shown in Algorithm 2, except that, in row 9, the statement **if** $(p \leqslant \frac{k_v}{k_w})$ is replaced by the statement **if** (w is a bridge) **or** $(p \leqslant \frac{k_v}{k_w})$. Observe that, in this case, the higher the degree of a non-bridge, the higher the probability that M^B discards it. This property does not hold for bridges who are always selected by M^B.
- M^{B_1}: Analogously to M^B, it has been conceived to correct M by favoring bridges. However, the effects of M^{B_1} on bridges are different from the ones of M^B. In words, the preference in favor of bridges is less drastic than M^B (see below). M^{B_1} starts by randomly generating two real numbers a and b belonging to the interval $(0,1)$. Then it sets $p_1 = min(a,b)$ and $p_2 = max(a,b)$. During each iteration it considers the currently visited node v and a node w randomly selected from the neighbors of v. If w is a bridge, then it moves from v to w if $p_1 \leqslant \frac{k_v}{k_w}$. Otherwise, if w is not a bridge, then it moves from v to w if $p_2 \leqslant \frac{k_v}{k_w}$. In all the other cases it stays in v. Since $p_1 \leqslant p_2$, bridges are favored. It terminates after n_{it} iterations. The pseudocode of this algorithm is shown in Algorithm 3. Clearly, if we force p_1 to 0, then M^{B_1} reduces to M^B. Observe that M^{B_1} is more selective on bridges than M^B, since bridges are visited only if the degree condition (i.e., $p_1 \leqslant \frac{k_v}{k_w}$) is verified. On the contrary, M^{B_1} is less selective on bridges than M (in the sense that the probability of selecting a high-degree bridge is higher than in M). This happens because the probability distribution of p_1, differently from the uniform distribution of p in M, is biased towards 0. Indeed, it is easy to see that the probability density function $f_{p_1}(x)$, where $x \in (0,1)$, of the random variable p_1 – i.e., the minimum between two real numbers randomly extracted in the interval $(0,1)$ – is $f_{p_1}(x) = 2 \cdot (1-x)$ (linearly decreasing[2]) against the density function $f_p(x) = 1$ (constant). Moreover, M^{B_1} is more selective also on non-bridges than M^B (and therefore M), since the degree condition (i.e., $p_2 \leqslant \frac{k_v}{k_w}$) operates with a bound p_2 whose probability distribution, differently from the uniform distribution of p in M^B and M, is biased towards 1. Indeed, it is easy to see that the probability density function $f_{p_2}(x)$, where $x \in (0,1)$, of the random variable p_2 – i.e., the maximum between two real numbers extracted randomly in the interval $(0,1)$ – is $f_{p_2}(x) = 2 \cdot x$ (linearly increasing) against the density function $f_p(x) = 1$ (constant).

[2] Indeed, the density function of the minimum between two real numbers R_1 and R_2 randomly extracted in a range $(0,a)$ is $f_{min}(x) = P(R_1 \geqslant x) \cdot \frac{1}{a} + P(R_2 \geqslant x) \cdot \frac{1}{a} = 2 \cdot \frac{1}{a} \cdot (1 - \int_0^x \frac{1}{a} dx) = \frac{2}{a^2} \cdot (a - x)$, whereas the density function of the maximum is $f_{max}(x) = P(R_1 \leqslant x) \cdot \frac{1}{a} + P(R_2 \leqslant x) \cdot \frac{1}{a} = 2 \cdot \frac{1}{a} \cdot \int_0^x \frac{1}{a} dx = \frac{2x}{a^2}$.

Algorithm 3. M^{B_1}

Input s: a seed node
Output *SeenNode*, *VisitedNode*: a set of nodes
Variable n_{it} {The number of iterations}
Variable v, w: a node
Variable p_1, p_2: a number in the real interval $(0,1)$
Variable a, b: a number in the real interval $(0,1)$

```
1: SeenNode := ∅, VisitedNode := ∅
2: insert s into SeenNode and VisitedNode
3: insert all the nodes adjacent to s into SeenNode
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:    randomly generate two numbers a and b in the real interval (0,1)
8:    p_1 = min(a, b)
9:    p_2 = max(a, b)
10:   let k_v and k_w be the outdegree of v and w, respectively
11:   if (w is a bridge) and (p_1 ⩽ k_v/k_w) then
12:      v = w
13:      insert w into VisitedNode
14:      insert all the nodes adjacent to w into SeenNode
15:   else
16:      if (v is not a bridge) and (p_2 ⩽ k_v/k_w) then
17:         v = w
18:         insert w into VisitedNode
19:         insert all the nodes adjacent to w into SeenNode
20:      end if
21:   end if
22: end for
```

Observe that *BFS* and *M* are classic crawling strategies, whereas BFS^B, M^B and M^{B_1} have been introduced in this paper in order to take the specific context into account.

A crawling strategy generally uses one or more accounts as seeds. In order to obtain samples as variegated and representative as possible, for each sample to obtain, we run the corresponding crawler from 10 different seeds and, then, merged data crawled from each seed. Moreover, in the definition of the starting seeds, we considered two options. In a first one, we chose seeds in a totally random fashion. In a second one, we chose seeds randomly, but only among bridges. As a consequence, we have defined the following 10 crawling running modes:

- *BFS*, BFS^B, M, M^B and M^{B_1}: in these running modes we have applied the *BFS*, BFS^B, M, M^B and M^{B_1} strategies and we have chosen seeds in a random way, among all nodes;
- $\overline{BFS}$, $\overline{BFS^B}$, $\overline{M}$, $\overline{M^B}$ and $\overline{M^{B_1}}$: in these running modes we have applied the *BFS*, BFS^B, M, M^B and M^{B_1} strategies and we have chosen seeds randomly, but only among bridges.

Since, to the best of our knowledge, our analysis is the first one specifically conceived for investigating SIS's, in order to better "dominate" (and, therefore, understand) the context of interest, we limited our crawlers to extract samples from only five OSNs, among the ones compliant with the XFN and FOAF standards. Specifically, selected OSNs are: Twitter, LiveJournal, YouTube, MySpace and Google+. We chose Twitter, LiveJournal and YouTube because they have been largely analyzed in the past in Social Network Analysis papers devoted to study a single OSN or to compare different OSNs. We selected MySpace because, differently from the other OSNs into consideration, it presents symmetrical connections among its accounts and we fstate the following implication, answering Question. Finally, we selected Google+ because it is a recent OSN and we judged that an analysis involving this OSN could have been very interesting for the Social Network Analysis research field.

3.2. Collected data

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 September 5, 2011 to October 20, 2011. For each running mode we run the corresponding crawler and we performed 15,000 iterations. In this way, we obtained 10 samples.

Table 1
Basic statistics of collected data (BFS, BFS^B, M, M^B and M^{B_1} running modes.).

	BFS	BFS^B	M	M^B	M^{B_1}
Seen nodes					
Total	3,523,588	6,572,238	949,181	6,178,649	312,598
Twitter	1,378,727	5,463,220	734,648	3,239,278	195,774
YouTube	209,858	178,828	31,729	258,618	13,868
MySpace	1,362,129	502,413	95,928	2,562,037	52,045
LiveJournal	147,100	33,170	51,923	4694	29,487
Google+	425,775	394,609	34,953	114,023	21,424
Visited nodes					
Total	13,605	10,978	10,933	12,565	6997
Twitter	2909	5072	5,265	5,113	3162
YouTube	2544	222	851	498	531
MySpace	2998	1943	3171	6220	2389
LiveJournal	1891	231	1090	164	688
Google+	3263	3510	556	570	227
Links	9,630,958	12,248,937	1,307,484	12,206,992	443,967
Me Links	3043	4779	526	4138	1150

Table 2
Basic statistics of the collected data ($\overline{BFS}$, $\overline{BFS^B}$, $\overline{M}$, $\overline{M^B}$ and $\overline{M^{B_1}}$ running modes).

	$\overline{BFS}$	$\overline{BFS^B}$	$\overline{M}$	$\overline{M^B}$	$\overline{M^{B_1}}$
Seen nodes					
Total	3,476,795	10,559,292	952,693	5,569,483	345,709
Twitter	1,508,274	8,367,007	738,939	2,368,777	210,514
YouTube	275,890	401,012	6721	282,122	8751
MySpace	925,227	1,395,272	158,612	2,767,507	112,831
LiveJournal	200,511	48,880	31,204	39,116	8536
Google+	566,895	347,121	17,218	111,962	5077
Visited nodes					
Total	14,676	11,165	11,808	11,727	7546
Twitter	3128	5588	4746	3793	2319
YouTube	2689	367	239	405	559
MySpace	3001	2852	5414	7166	3728
LiveJournal	2780	231	633	139	520
Google+	3078	2127	776	224	420
Links	10,208,799	20,133,682	1,320,667	11,569,695	516,971
Me Links	5124	4777	436	3657	496

These samples can be found at the URL address http://www.ursino.unirc.it/bridges.html. For each sample we preliminarily measured the following basic statistics, i.e.:

- the number of seen nodes and visited nodes[3];
- the number of links;
- the number of `me` links; and
- the distributions of seen nodes and visited nodes for the five considered OSNs.

All these statistics are reported in Tables 1 and 2.

4. Bridge analysis

In this section, we describe our experimental analysis aiming at answering the questions stated in the introduction. Each sub-section provides the answer to one or more questions given in the introduction. Furthermore, in the last sub-section we report an interesting additional result whose scope is related to the OSNs chosen in our SIS.

4.1. Bridges and backbones

In this analysis, we aim at verifying whether there is a sort of "backbone" among bridges, favoring the direct links among them. To do this, we analyze the percentage of bridges in the samples returned by the crawling tasks. We recall that, for each

[3] Visited nodes are also called crawled nodes in the literature [64].

Table 3
Percentage of bridges in each crawled sample.

	BFS	$\overline{BFS}$	BFS^B	$\overline{BFS^B}$	M	$\overline{M}$	M^B	$\overline{M^B}$	M^{B_1}	$\overline{M^{B_1}}$
Bridge (%)	15.83	19.45	23.21	23.25	4.34	3.52	14.91	13.18	7.96	5.98

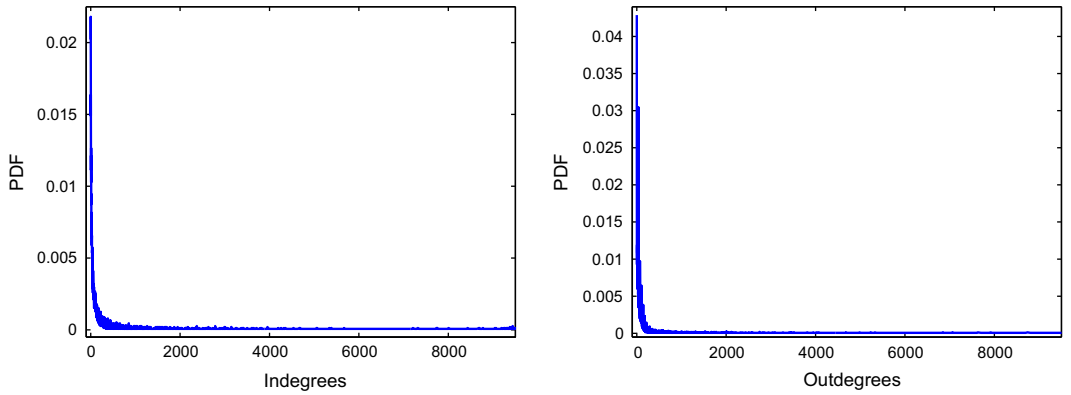

Fig. 1. Probability distribution function of bridge degrees.

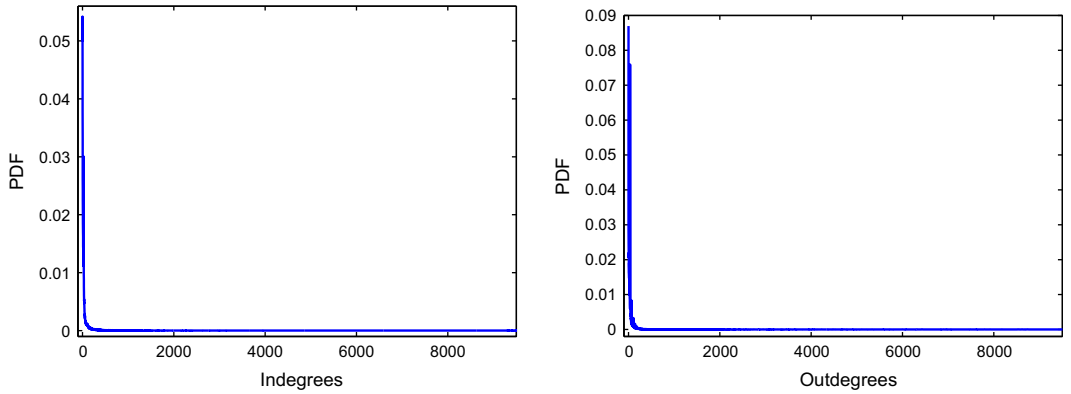

Fig. 2. Probability distribution function of node degrees.

crawling strategy, we consider two running modes. In the first one we choose seed nodes in a random way among all nodes, whereas in the second one we choose seed nodes randomly, but only among bridges. We argued that, if a "backbone" exists among bridges, a sample produced by a running mode starting from bridges (i.e., $\overline{BFS}$, BFS^B, $\overline{M}$, M^B and M^{B_1}) should have a percentage of bridges higher than the one occurring in the sample obtained by the same crawling strategy started from seeds selected among all nodes. The percentages of bridges occurring in each crawled sample are reported in Table 3.

From the analysis of this table we can see that, given a crawling strategy, the percentage of bridges in the corresponding samples does not significantly vary whenever the seed nodes are chosen among all nodes or they are selected only from bridges. This result allows us to state the following implication, answering Question 1 in the introduction:

Implication 1. In a SIS, it does not exist a sort of "backbone" among bridges, aiming at favoring the direct links among them.

4.2. Indegree and outdegree distributions of bridges

The purpose of this analysis is to investigate the distributions of the indegree and the outdegree of bridges, comparing them with the corresponding ones of nodes. In order to maximize the size of the sample exploited in this investigation, we consider a unique sample obtained by the union of the ten samples considered in the other experiments. As a first step we analyze the distributions of the indegree and the outdegree of the bridges and the nodes in the sample. The corresponding Probability Distribution Functions (PDFs) are shown in Figs. 1 and 2, respectively.

Table 4
Power law coefficient estimates and corresponding Kolmogorov–Smirnov goodness-of-fit metrics for indegrees and outdegrees of nodes and bridges.

	Indegrees		Outdegrees	
	α	D	α	D
Nodes	1.64	0.127	1.58	0.144
Bridges	2.23	0.147	2.92	0.159

A simple visual analysis of these diagrams leads us to guess that the indegree and the outdegree of both bridges and nodes follow a power law distribution. In order to verify this conjecture we compute the best power law fit using the maximum likelihood method [23]. Table 4 shows the estimated power law coefficient, along with the Kolmogorov–Smirnov goodness-of-fit metric [23], for the four distributions into consideration.

From the analysis of this table it is evident that both the indegree and the outdegree of the bridges and the nodes of a SIS follow a power law distribution very well. In fact, in all the considered cases, the values of the Kolmogorov–Smirnov goodness-of-fit metric are low [49]. Interestingly, the power law coefficients of bridges are higher than (even though close to) the ones of nodes for both the indegree and the outdegree distributions. In sum, we may conclude that:

Implication 2. In a SIS, the indegree and the outdegree of bridges follow a power law distribution.

This implication, along with the results about the power law coefficients of bridges and non-bridges, is very important not only because it answers Question 2 of the introduction, but also because it allows us to conjecture an answer to Question 3. Indeed, from a theoretical point of view, Implication 2 should, in its turn, imply that an increase of the fraction of bridges in an OSN should lead to an increase of the average indegree and outdegree of nodes as well as of the standard deviation of node degree. We perform an experiment to verify this conjecture. In particular, we compute the average indegree and outdegree of nodes,[4] along with the corresponding standard deviations, for each OSN of the SIS. Obtained results are shown in Table 5. In order to facilitate the analysis, in this table we report also the fraction of bridges for each crawled sample, as already shown in Table 3.

From the analysis of this table, it is possible to observe that, in a SIS, an increase of the fraction of bridges generally leads to an increase of the average indegree and outdegree of nodes, along with the corresponding standard deviations, for each OSN. This reasoning allows us to state the following corollary:

Corollary 2.1. *In a SIS, the higher the fraction of bridges, the higher the average indegree and outdegree of nodes, and the higher the corresponding standard deviations of the associated OSNs.*

Therefore, the above conjecture holds. Consider that, in presence of a power law distribution, the concepts of average indegree and outdegree of bridges and non-bridges are little significant as absolute values. However, they can become interesting if compared with each other. With regard to this aspect, from a theoretical point of view, Corollary 2.1 implies a further one, which allows us to conjecture an answer to Question 4 of the introduction. In fact, Corollary 2.1 implies, in its turn, that the average indegree and outdegree of bridges are generally higher than the corresponding ones of non-bridges. In order to experimentally verify this last conjecture, for each crawled sample, we compute the average indegree and outdegree of the bridges and non-bridges for each OSN of the sample. Obtained results are shown in Table 6.

By comparing indegrees and outdegrees of bridges with the ones of non-bridges, it immediately follows the next corollary:

Corollary 2.2. *In a SIS, the average indegree and outdegree of bridges are generally higher than the corresponding ones of non-bridges.*

Thus, the latter conjecture holds too. In sum, Implication 2 and its corollaries answer Questions 2, 3 and 4 of the introduction.

4.3. Relationship between bridges and power users

Power users are those nodes which can facilitate information spread in the network. Node indegree and outdegree are considered good indicators to detect whether a node is a power user. Our notion of power user is indeed based on these indicators. In fact, for power users, we use the notion adopted in [1] (one of the most famous libraries which provides features for OSN/graph analysis). According to this notion, a power user is a node of an OSN with a degree higher than the average degree of the nodes (see below).

Implication 2, along with its corollaries, states that the presence of bridges leads to an increment of the average degree of nodes in a SIS and that the average degree of bridges is higher than the one of non-bridges. If the degree probability distri-

[4] It is worth pointing out that, in this computation, we analyze the indegrees and the outdegrees of the nodes by considering all the links declared by the corresponding users through the XFN and FOAF standards.

Table 5
Average indegree and outdegree of nodes, and corresponding standard deviations, for each crawling strategy and each OSN into examination.

	In degree		Out degree	
	Avg	Std	Avg	Std
BFS (15.83%)				
Total	350.64	1290.59	367.97	1289.37
Twitter	482.20	1559.84	389.67	769.04
YouTube	79.34	693.26	11.84	9.65
MySpace	623.32	2021.66	627.14	2035.16
LiveJournal	67.34	306.98	81.93	155.06
Google+	358.54	635.14	553.93	1530.32
$\overline{BFS}$ *(19.45%)*				
Total	305.30	1163.13	405.54	1411.24
Twitter	504.28	1574.20	380.21	895.82
YouTube	102.63	814.19	12.13	11.08
MySpace	375.20	1593.61	376.75	1600.36
LiveJournal	65.05	290.74	70.67	145.98
Google+	428.98	807.65	1105.48	2331.56
BFS^B *(23.21%)*				
Total	848.04	2151.52	273.26	1180.54
Twitter	1529.63	2796.69	206.33	832.36
YouTube	824.98	1999.51	10.08	11.66
MySpace	299.38	1512.14	300.96	1520.70
LiveJournal	135.77	733.87	52.80	141.08
Google+	215.18	591.53	385.79	1431.50
$\overline{BFS^B}$ *(23.25%)*				
Total	1430.64	4082.92	376.79	1976.51
Twitter	2341.95	5124.69	189.01	1171.34
YouTube	1155.83	3506.64	11.16	11.15
MySpace	642.51	2911.26	647.85	2936.84
LiveJournal	237.96	818.94	94.82	179.32
Google+	270.20	649.03	600.38	2249.57
M (4.34%)				
Total	69.00	503.61	52.30	302.91
Twitter	105.51	694.61	69.19	354.01
YouTube	32.68	290.52	9.41	8.63
MySpace	34.37	187.74	34.48	188.92
LiveJournal	37.29	115.12	43.59	110.33
Google+	38.63	194.09	76.67	617.96
$\overline{M}$ *(3.52%)*				
Total	65.75	492.24	47.94	250.76
Twitter	115.80	743.57	71.64	333.10
YouTube	22.95	120.04	9.70	8.17
MySpace	33.31	193.75	33.36	194.42
LiveJournal	36.04	116.00	39.20	78.44
Google+	23.37	49.72	23.59	58.52
M^B *(14.91%)*				
Total	607.51	2125.66	366.02	1597.44
Twitter	772.17	2381.85	206.26	958.61
YouTube	531.28	1805.97	10.12	9.05
MySpace	535.04	2030.66	536.68	2037.16
LiveJournal	16.21	27.61	27.04	48.93
Google+	158.03	507.93	345.22	1424.99
$\overline{M^B}$ *(13.18%)*				
Total	598.02	2127.28	390.57	1679.13
Twitter	760.10	2396.83	172.33	854.28
YouTube	710.49	2223.98	10.69	10.06
MySpace	518.44	2001.16	519.84	2006.86
LiveJournal	265.12	1243.22	94.22	190.19
Google+	402.52	918.70	821.21	2130.91
M^{B_1} *(7.96%)*				
Total	34.12	283.58	31.03	149.38
Twitter	42.18	398.70	34.04	94.05
YouTube	22.54	156.01	8.28	7.93
MySpace	25.47	65.01	25.47	65.17
LiveJournal	31.36	195.95	31.92	80.58

(*continued on next page*)

Table 5 (*continued*)

	In degree		Out degree	
	Avg	Std	Avg	Std
Google+	48.37	210.78	98.06	702.78
$\overline{M}^{B_1}$ *(5.98%)*				
Total	39.27	401.95	31.10	258.32
Twitter	65.64	606.87	39.99	248.36
YouTube	12.38	75.75	7.50	7.07
MySpace	33.94	310.07	33.95	310.35
LiveJournal	11.71	17.46	13.81	24.81
Google+	10.82	32.32	9.51	22.88

bution were not taken into consideration, one would erroneously conclude that it is probable to find bridges with high degree, thus, that bridges are typically power users. But, thanks to Implication 2, we know that the degree of bridges follows a power law distribution. This allows us to exclude the above conclusion, since the power law distribution means that only a few bridges have a very high degree, whereas most of bridges have a low degree. In other words, the effect of increasing the average degree for bridges depends on just a few bridges, where the high degree is concentrated.

The question now is whether the power users of the SIS are just those bridges with high degree and vice versa. In other words, we have to understand how much the set of power users and the one of bridges are overlapping. We distinguish between *in power users* and *out power users*, depending on which degree we consider between indegree and outdegree, respectively. In particular, we denote by *InPU* (*OutPU*, resp.) the set of nodes of the sample having an indegree (outdegree, resp.) higher than the average indegree (outdegree, resp.) of the nodes of the sample. We also denote by B the set of bridges of the sample.

From the reasoning above, we expect that both $InPB = \frac{|InPU \cap B|}{|B|}$ and $OutPB = \frac{|OutPU \cap B|}{|B|}$ (expressing the fraction of bridges who are also power users) are low. However, it could happen that $InBP = \frac{|InPU \cap B|}{|InPU|}$ or $OutBP = \frac{|OutPU \cap B|}{|OutPU|}$ (expressing the fraction of power users who are also bridges) is high. This would mean that most of the power users are bridges. The experiment described in this section aims at confirming the expectation about *InPB* and *OutPB* and to study *InBP* and *OutBP*.

The results of the experiment are shown in Table 7. Here, it is possible to see that, in general, all the coefficients are low. On the one hand, this confirms our expectation that the probability that a bridge is a power user is low (basically due to Implication 2). On the other hand, the experiment allows us to discover that even the probability that a power user is a bridge is low. In sum, we may conclude that there does not exist a meaningful correlation between bridges and power users. The only exception regards the crawling strategy M^B. This can be explained by considering the philosophy underlying this strategy, which tends to penalize power users except when they are bridges (since bridges are always selected in this strategy – see Section 3.1). As a consequence, most of the power users selected by this strategy are bridges. Even for M^{B_1} we obtained values slightly higher than the other techniques, but less than M^B. This can be explained by recalling that also M^{B_1} is biased towards bridges, but less than M^B (see, again, Section 3.1).

In order to complete the study of the correlation between power users and bridges, we come back to a previous consideration derived from Implication 2. We said above that the bridge effect of increasing the average degree, derived from Implication 2 and its corollaries, depends on just a few bridges, where the high degree is concentrated. This is, in fact, what we have verified through our experiment. At this point it becomes interesting to understand "how much" these few bridges with high degree are power users. In other words, we want to study whether a correlation between bridges and power users emerges if we "stress" the definition of power user. In particular, we computed the above coefficients *InPB*, *OutPB*, *InBP*, and *OutBP* by considering the sets $InPU_x$ and $OutPU_x$, representing the top x% of power users having the highest indegree and outdegree, respectively. x ranges from 10 to 100 with granularity 10. We denote by $InPB_x$, $OutPB_x$, $InBP_x$, and $OutBP_x$ the four coefficients. Obviously, $InPU_{100}$ reduces to *InPU* and $OutPU_{100}$ to *OutPU*. As a consequence, $InPB_{100}$, $OutPB_{100}$, $InBP_{100}$ and $OutBP_{100}$ coincide with *InPB*, *OutPB*, *InBP*, and *OutBP*, respectively. The lower the value of x, the "stronger" (in terms of degree) the considered power users. We want to understand whether a correlation between power users and bridges arises by increasing the strongness of power users. To this aim, we are interested in studying how the four coefficients vary when x increases. Indeed, we expect that if no correlation between bridges and power users exists (also for increasing strongness), we should obtain:

- A quasi-linear dependence of the coefficients $InPB_x$ and $OutPB_x$ on the value of x, showing that the intersection between bridges and power users decreases proportionally in the same measure as x reduces the cardinality of the power user sets (observe that the power user sets $InPU_x$ and $OutPU_x$ occur only in the numerator of the respective coefficients).
- A quasi-constant behavior of the coefficients $InBP_x$ and $OutBP_x$ when the value of x varies. It is due to the same reasons of the previous case, but considering that the power user sets $InPU_x$ and $OutPU_x$ occur in both the numerator and the denominator of the respective coefficients.

Table 6
Average indegree and outdegree of bridges and non-bridges for each crawled sample.

	Bridges		Non-bridges	
	Avg IN	Avg OUT	Avg IN	Avg OUT
BFS				
Total	764.83	700.76	272.73	305.37
Twitter	901.43	596.95	333.41	316.10
YouTube	232.81	15.94	64.67	11.45
MySpace	2626.54	2641.38	485.48	488.55
LiveJournal	195.89	190.50	48.30	65.84
Google+	483.14	674.98	322.44	518.86
$\overline{BFS}$				
Total	617.26	816.53	229.99	306.32
Twitter	974.48	721.28	396.57	302.08
YouTube	340.80	16.24	73.17	11.62
MySpace	2977.45	2990.88	270.57	271.64
LiveJournal	108.26	97.15	52.10	62.74
Google+	556.56	1228.12	345.44	1025.17
BFS^B				
Total	1037.25	396.48	790.85	236.01
Twitter	1877.54	247.26	1435.95	195.30
YouTube	587.78	8.77	950.94	10.78
MySpace	3348.12	3352.07	207.24	208.75
LiveJournal	529.05	118.71	58.34	39.82
Google+	281.90	421.48	175.93	364.80
$\overline{BFS^B}$				
Total	2058.88	556.01	1240.32	322.50
Twitter	2932.03	214.57	2132.28	179.93
YouTube	1401.62	11.78	1031.92	10.85
MySpace	4752.57	4768.38	455.69	460.56
LiveJournal	275.11	89.68	229.22	96.03
Google+	328.11	633.80	232.40	578.58
M				
Total	238.84	104.88	61.29	49.91
Twitter	420.25	126.09	92.24	66.79
YouTube	134.22	12.22	25.66	9.21
MySpace	78.71	78.09	33.58	33.71
LiveJournal	92.56	101.07	32.61	38.73
Google+	65.06	140.34	35.01	67.94
$\overline{M}$				
Total	234.71	79.32	59.58	46.79
Twitter	456.34	100.60	102.99	70.55
YouTube	102.27	8.20	11.56	9.92
MySpace	105.25	104.30	32.22	32.28
LiveJournal	69.49	61.49	32.08	36.56
Google+	43.92	43.67	21.46	21.73
M^B				
Total	3604.46	2043.39	82.19	72.00
Twitter	4039.48	711.42	138.10	108.22
YouTube	2275.51	12.19	13.46	9.51
MySpace	4024.60	4036.31	47.13	47.36
LiveJournal	33.89	75.58	13.90	20.68
Google+	456.59	1058.42	54.27	97.37
$\overline{M^B}$				
Total	4045.37	2519.68	74.53	67.26
Twitter	4400.10	570.64	125.64	102.90
YouTube	2956.74	11.50	31.56	10.44
MySpace	4369.85	4381.79	50.04	50.16
LiveJournal	1100.21	267.57	54.47	50.50
Google+	866.26	1915.86	124.28	164.43
M^{B_1}				
Total	199.71	102.69	19.80	24.83
Twitter	318.64	105.85	20.19	28.33
YouTube	121.15	13.89	6.32	7.36
MySpace	117.11	116.38	22.21	22.24
LiveJournal	100.72	89.79	19.56	22.08

(*continued on next page*)

Table 6 (*continued*)

	Bridges		Non-bridges	
	Avg IN	Avg OUT	Avg IN	Avg OUT
Google+	122.93	193.61	17.16	58.04
$\overline{M^{B_1}}$				
Total	340.40	172.86	20.13	22.09
Twitter	687.37	202.05	23.56	29.02
YouTube	64.06	10.83	5.10	7.04
MySpace	388.03	387.50	22.67	22.70
LiveJournal	31.15	33.46	9.82	11.90
Google+	27.05	19.03	7.35	7.47

Table 7
Values of *InPB*, *OutPB*, *InBP* and *OutBP* for each crawled sample.

	InPB	OutPB	InBP	OutBP
BFS	0.268	0.273	0.333	0.310
$\overline{BFS}$	0.306	0.248	0.446	0.370
BFS^B	0.208	0.138	0.288	0.337
$\overline{BFS}^B$	0.248	0.094	0.365	0.349
M	0.248	0.314	0.097	0.089
$\overline{M}$	0.281	0.349	0.089	0.074
M^B	0.519	0.353	0.848	0.701
$\overline{M^B}$	0.552	0.391	0.854	0.720
$\mathbf{M}^{B_1}$	0.427	0.557	0.240	0.159
$\overline{M^{B_1}}$	0.408	0.503	0.274	0.141

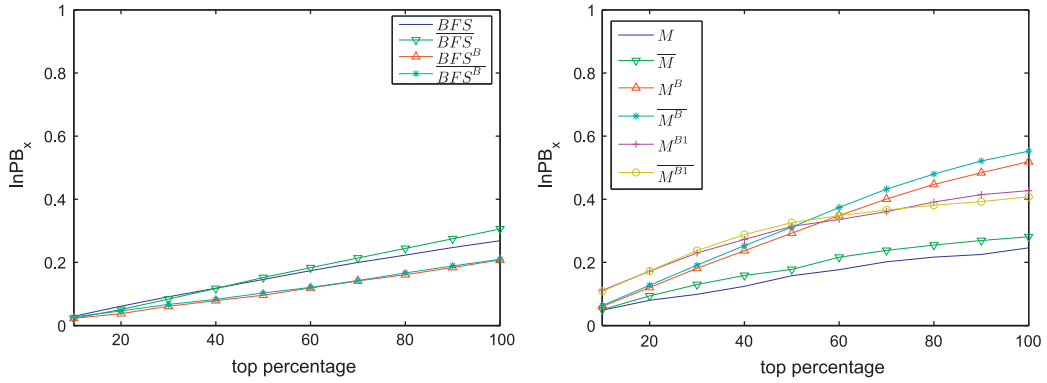

Fig. 3. $InPB_x$ vs top percentage.

As it can be derived by analyzing Figs. 3–6, experimental results show that we are in the scenario above hypothesized. The only exception regards the crawling strategies M^B and M^{B_1}. The biased behavior of these strategies is more evident for the coefficients $InBP_x$ and $OutBP_x$, where it is amplified by the reduction of the denominator, which is more rapid than the one of the numerator – observe that the denominator $|InPU_x|$ ($|OutPU_x|$, resp.) of $InBP_x$ and ($OutBP_x$, resp.) decreases as x decreases since the number of considered power users decreases. The bias detected for M^B and M^{B_1} (more evident for M^{B_1}) can be explained by considering how these strategies work. About M^B, we have observed in Section 3.1 that the higher the degree of a non-bridge, the higher the probability that M^B discards it. As a consequence, the stronger (in terms of degree) a non-bridge, the lower the probability it belongs to the set occurring in the numerator of the coefficients. This works in favor of the increase of the proportion of bridges belonging to the sets occurring in the coefficient numerators, thus contrasting their decrease (as x decreases). This explains why, for decreasing values of x (i.e., increasing strongness), the coefficients $InBP_x$ and $OutBP_x$ are biased towards higher values w.r.t. the expected behavior (i.e., constant). This effect is more evident in M^{B_1} since, as observed in Section 3.1, this strategy is more selective than M^B on non-bridges nodes. From all the above reasonings the following implication arises, answering Question 5 of the introduction:

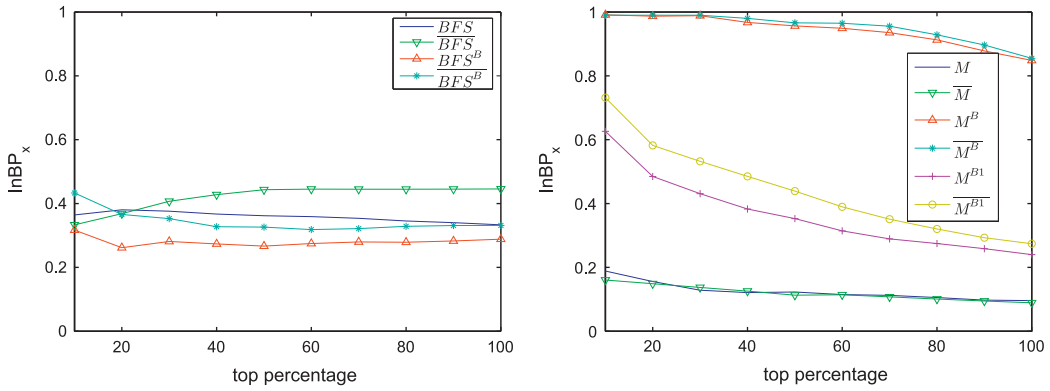

Fig. 4. *$InBP_x$* vs top percentage.

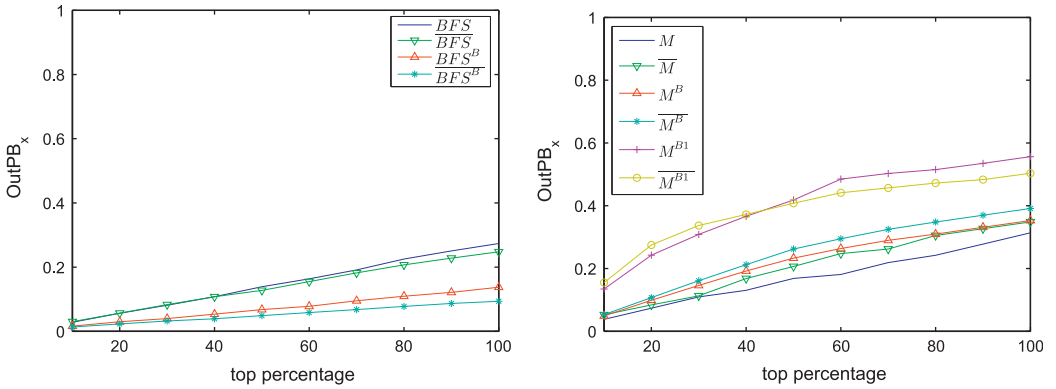

Fig. 5. *$OutPB_x$* vs top percentage.

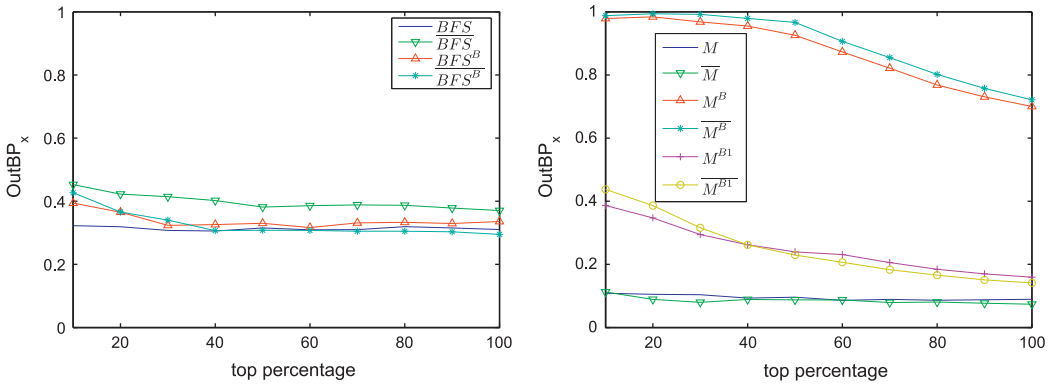

Fig. 6. *$OutBP_x$* vs top percentage.

Implication 3. In a SIS, there is no correlation between bridges and power users.

This result is important from the structural point of view due to the role of bridges in the interconnection of the OSNs of the SIS. Moreover, it could provide interesting hints for the analysis of user behavior, which is anyway out of the scope of this work.

4.4. Capability of bridges of connecting more OSNs

In this experiment we measure the number of OSNs connected by the available bridges in each crawled sample. Observe that, since our SIS consists of 5 OSNs, the number of OSNs that can be connected by a bridge ranges between 2 and 5. The

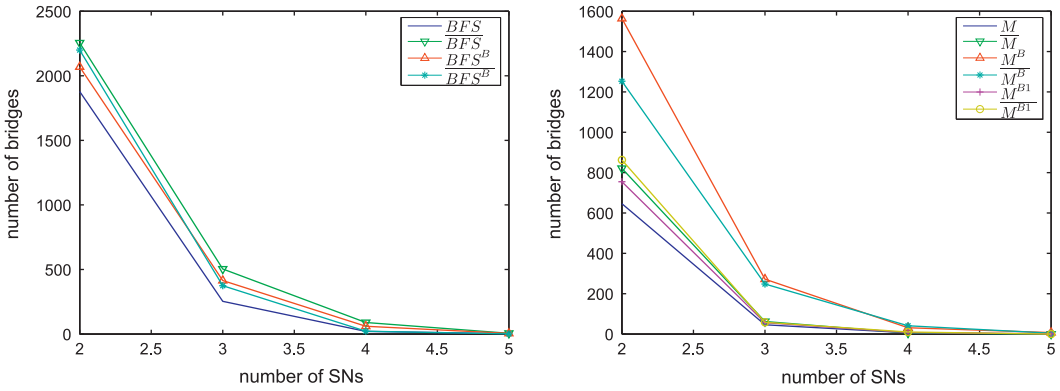

Fig. 7. Number of OSNs connected by the available bridges.

Table 8
Bridge percentage in the SIS and in the corresponding OSNs.

	SIS (%)	Twitter (%)	YouTube (%)	MySpace (%)	LiveJournal (%)	Google+ (%)
BFS	15.83	26.19	8.73	6.44	12.90	22.46
$\overline{BFS}$	19.45	18.64	11.01	3.87	23.06	39.57
BFS^B	23.21	21.21	34.68	2.93	16.45	37.04
$\overline{BFS^{\mathbf{B}}}$	23.25	26.22	33.51	4.35	19.05	39.49
M	4.34	4.05	6.46	1.73	7.80	12.05
$\overline{M}$	3.52	3.62	12.55	1.50	10.58	8.51
M^B	14.91	16.25	22.89	12.27	11.59	25.79
$\overline{M^B}$	13.18	14.84	23.21	10.84	20.14	37.50
$\mathbf{M}^{B_1}$	7.96	7.37	14.12	3.43	14.53	29.52
$\overline{M^{B_1}}$	5.98	6.34	12.34	3.08	8.85	17.62

results of this experiment are shown in Fig. 7. From the analysis of this figure it is possible to observe that, for each crawling strategy, most bridges connect a few OSNs and a few bridges connect many OSNs or all of them. For instance, if we consider all crawled samples, then the percentage of bridges which connect only 2 OSNs ranges from 79.01% to 92.70%, whereas the fraction of bridges which connect 5 OSNs ranges from 0% to 2.25%. The trends shown in the previous figure qualitatively look like a power law distribution. However, the number of OSNs composing our SIS is small. As a consequence, the number of values considered in the axes of the previous figure is small too. This fact makes little meaningful to quantitatively verify whether the number of OSNs connected by bridges follows a power law distribution and, in this case, to compute the corresponding α and D coefficients. On the other hand, the huge volume of analyses performed in our experiments made it prohibitive to extend the number of considered OSNs beyond 5, besides the reasons expressed in Section 3.1 for this choice. Therefore, in the wide-spectrum analysis on bridges faced in this paper, we considered satisfactory a qualitative analysis of this aspect, delaying a quantitative one to a specific future work where taking all the OSNs handling XFN and FOAF into account is feasible. The conclusion we can draw from the above analysis can be thus summarized by the following implication, answering Question 6 of the introduction:

Implication 4. In a SIS, most of the bridges connect a few OSNs, whereas a few bridges connect many OSNs.

4.5. Fraction of bridges

In this analysis we describe a result whose scope is related to the OSNs chosen in our SIS. In particular, for each crawled sample, we computed the percentage of nodes which are bridges in each OSN of the SIS. The results are shown in Table 8.

From the analysis of this table it emerges an interesting result. Indeed, it is possible to observe that, in Google+, the percentage of nodes which are bridges is generally higher than the one in the other OSNs. This non-obvious result can be explained by considering the characteristics of Google+. In fact, this OSN is quite recent and, thus, most of its users have already joined other OSNs. Furthermore, Google+ provides its users with very friendly utilities allowing them to specify their accounts in other OSNs and to import the corresponding data. All these facts favor the presence of bridges in Google+. This analysis leads us to claim the following implication:

Implication 5. Among the five OSNs of our SIS, Google+ generally presents the highest percentage of users who are bridges.

5. Conclusion and future work

This paper explores the emergent scenario of Social Internetworking from the perspective of Social Network Analysis. Being aware that the complete investigation of all the aspects of SIS's is an extremely large task, we have identified the most basic structural peculiarity of a SIS, i.e. bridges, and we have deeply studied it. We argue that most of the knowledge about the structural properties of SIS's, and possibly also about the behavioral aspects of users, starts from the adequate knowledge of bridges, which are the structural pillars of SIS's.

Our work was conducted with strong attention on the crawler strategy to adopt, through both the adaptation of existing techniques and the definition of new specific ones. This was also a chance for giving some initial results about the behavior of the various strategies in this particular context, preparing us for a further study on crawling a SIS. But the main reason of the application of a multiple-crawling-strategy approach was to guarantee an adequate scientific support to the implications we have derived about several structural features concerning bridges.

Let us summarize now these implications, and try to draw some conclusions. First, we have discovered that bridges, like power users, are nodes that contribute to increase the average degree of the SIS (Implication 2 and its corollaries), but they are not so "strong" like power users. Then, we have shown that there is no correlation between power users and bridges (Implication 3), and this analysis resulted in a number of very interesting aspects. We have seen that the results above are not in contradiction, because bridges follow a power law distribution, making very probable to have a low degree for a bridge (Implication 2) even though the average degree is higher than the one of non-bridges. A possible interpretation of the above knowledge is that bridges are users who are active in at least one OSN involved by their `me` links, so that they form a population of users not including a significant percentage of fake (or "one-time") users. Of course, many future investigations can be focused on this and other behavioral aspects. But coming back to the results of this papers, we have also seen that bridges do not form backbones among the involved OSNs (Implication 1), as one could intuitively expect. Finally, we have seen that most of the bridges connect a few OSNs, whereas a few bridges connect many OSNs, enabling a number of possible considerations from the behavioral point of view.

In conclusion, we think that this paper gives interesting results on the emergent scenario of Social Internetworking which open a number of interesting issues about both structural and behavioral aspects. We plan to move our future research towards these directions, including a deep study on the crawling of SIS's, and possibly applying Data Warehousing and Data Mining techniques on crawled samples.

References

[1] Stanford Network Analysis Package, 2011. <http://snap.stanford.edu/snap/>
[2] FriendFeed, 2012. <http://friendfeed.com/>
[3] Gathera, 2012. <http://www.gathera.com/>
[4] Google Open Social, 2012. <http://code.google.com/intl/it-IT/apis/opensocial/>
[5] Google Social Graph, 2012. <http://code.google.com/p/itswhoyouknow/wiki/SocialGraph>
[6] Power.com, 2012. <http://power.com>
[7] The Friend of a Friend (FOAF) Project, 2012. <http://www.foaf-project.org/>
[8] XFN – XHTML Friends Network, 2012. <http://gmpg.org/xfn>
[9] F. Abel, N. Henze, E. Herder, D. Krause, Interweaving public user profiles on the web, in: Proc. of the International Conference on User Modeling, Adaptation, Personalization (UMAP'10), Big Island, Hawaii, USA, Lecture Notes in Computer Science, Springer, 2010, pp. 16–27.
[10] N. Agarwal, M. Galan, H. Liu, S. Subramanya, WisColl: collective wisdom based blog clustering, Information Sciences 180 (1) (2010) 39–61.
[11] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge online social networking services, in: Proc. of the International Conference on World Wide Web (WWW'07), ACM, Banff, Alberta, Canada, 2007, pp. 835–844,.
[12] E.M. Airoldi, K.M. Carley, Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings, ACM SIGKDD Explorations Newsletter 7 (2) (2005) 13–22.
[13] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, and evolution, in: Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), ACM, Philadelphia, USA, 2006, pp. 44–54.
[14] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi, Foundations of multidimensional network analysis, in: Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011), IEEE, Kaohsiung, Taiwan, 2011, pp. 485–489.
[15] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi, The pursuit of hubbiness: analysis of hubs in large multidimensional networks, Journal of Computational Science 2 (3) (2011) 223–237.
[16] J. Bonneau, J. Anderson, G. Danezis, Prying data out of a social network, in: Proc. of the International Conference on Advances in Social Network Analysis and Mining (ASONAM'09), IEEE, Athens, Greece, 2009, pp. 249–254.
[17] M. Bröcheler, A. Pugliese, V.S. Subrahmanian, subgraph matching on huge social networks, in: Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011), IEEE, Kaohsiung, Taiwan, 2011, pp. 271–278.
[18] F. Buccafurri, G. Lax, A. Nocera, D. Ursino, Crawling social internetworking systems, in: Proc. of the International Conference on Advances in Social Analysis and Mining (ASONAM 2012), IEEE, Istanbul, Turkey, 2012, pp. 505–509.
[19] F. Buccafurri, G. Lax, A. Nocera, D. Ursino, Discovering links among social networks, in: Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012), Lecture Notes in Computer Science, Springer, Bristol, UK, 2012, pp. 467–482.
[20] P. Carrington, J. Scott, S. Wasserman, Models and Methods in Social Network Analysis, Cambridge University Press, 2005.
[21] S.A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Crawling Facebook for social network analysis purposes, in: Proc. of the International Conference Series on Web Intelligence, Mining and Semantics (WIMS'11), ACM, Sogndal, Norway, 2011, pp. 52–59.
[22] D.H. Chau, S. Pandit, S. Wang, C. Faloutsos, Parallel crawling for online social networks, in: Proc. of the International Conference on World Wide Web (WWW'07), ACM, Banff, Alberta, Canada, 2007, pp. 1283–1284.
[23] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-Law Distributions in Empirical Data, SIAM Review 51 (4) (2009) 661–703.
[24] M. Coscia, F. Giannotti, D. Pedreschi, A classification for community discovery methods in complex networks, Statistical Analysis and Data Mining 4 (5) (2011) 512–546.

[25] B.T. Dai, F.C.T. Chua, E.P. Lim, C. Faloutsos, Structural analysis in multi-relational social networks, in: Proc. of the International SIAM Conference on Data Mining (SDM 2012), Omnipress, Anaheim, CA, USA, 2012, pp. 451–462.
[26] P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Generalized Louvain method for community detection in large networks, in: Proc. of the International Conference on Intelligent Systems Design and Applications (ISDA 2011), IEEE, Cordoba, Spain, 2011, pp. 88–93.
[27] P. De Meo, A. Nocera, G. Terracina, D. Ursino, Recommendation of similar users, resources and social networks in a Social Internetworking Scenario, Information Sciences 181 (7) (2011) 1285–1305 (Elsevier).
[28] W. De Nooy, A. Mrvar, V. Batagelj, Exploratory Social Network Analysis with Pajek, Cambridge University Press, 2011.
[29] I. de Sola Pool, M. Kochen, Contacts and influence, Social Networks 1 (1978) 5–51.
[30] R. Ghosh, K. Lerman, Predicting influential users in online social networks, in: Proc. of the KDD International Workshop on Social Network Analysis (SNA-KDD'10), ACM, San Diego, CA, USA, 2010.
[31] C. Giatsidis, K. Berberich, D.M. Thilikos, M. Vazirgiannis, Visual exploration of collaboration networks based on graph degeneracy, in: Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012), ACM, Beijing, China, 2012, pp. 1512–1515.
[32] A.C. Gilbert, K. Levchenko, Compressing network graphs, in: Proc. of the International Workshop on Link Analysis and Group Detection (LinkKDD'04), ACM, Seattle, WA, USA, 2004.
[33] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: Proc. of the International Conference on Human Factors in Computing Systems (CHI'09), ACM, Boston, MA, USA, 2009, pp. 211–220.
[34] M. Gjoka, M. Kurant, C.T. Butts, A. Markopoulou, Walking in Facebook: A case study of unbiased sampling of OSNs, in: Proc. of the International Conference on Computer Communications (INFOCOM'10), IEEE, San Diego, CA, USA, 2010, pp. 1–9.
[35] M.S. Granovetter, The strength of weak ties, American Journal of Sociology 78 (6) (1973) 1360–1380.
[36] OFCOM The independent regulator and competition authority for the UK communications industries, Social Networking: A Quantitative and Qualitative Research Report into Attitudes, behaviours and use, 2009. <http://www.ofcom.org.uk/advice/medialiteracy/medlitpub/medlitpubrss/socialnetworking/annex3.pdf>
[37] J. Iturrioz, O. Diaz, C. Arellano. Towards federated Web2.0 sites: the TAGMAS approach, in: Proc. of the International Workshop on Tagging and Metadata for Social Information Organization, Banff, Alberta, Canada, 2007. <http://www2007.org/workshops/paper34.pdf>
[38] P. Kazienko, K. Musial, E. Kukla, T. Kajdanowicz, P. Bródka, Multidimensional social network: model and analysis, in: Proc. of the International Conference on Computational Collective Intelligence. Technologies and Applications (ICCCI 2011), Lecture Notes in Computer Science, Elsevier, Gdynia, Poland, 2011, pp. 378–387.
[39] J. Kleinberg, The convergence of social and technological networks, Communications of the ACM 51 (11) (2008) 66–72.
[40] A. Korolova, R. Motwani, S.U. Nabar, Y. Xu, Link privacy in social networks, in: Proc. of the ACM International Conference on Information and Knowledge Management (CIKM'08), ACM, Napa Valley, CA, USA, 2008, pp. 289–298.
[41] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.H. Cui, A. Percus, Reducing large internet topologies for faster simulations, in: Proc. of the International Conference on Networking (Networking 2005), Springer, Waterloo, Ontario, Canada, 2005, pp. 165–172.
[42] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, Link Mining: Models, Algorithms, and Applications (2010) 337–357.
[43] M. Kurant, A. Markopoulou, P. Thiran, On the bias of BFS (Breadth First Search), in: Proc. of the International Teletraffic Congress (ITC 22), IEEE, Amsterdam, The Netherlands, 2010, pp. 1–8.
[44] J. Leskovec, C. Faloutsos, Sampling from large graphs, in: Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), ACM, Philadelphia, PA, USA, 2006, pp. 631–636.
[45] Y.M. Li, C.Y. Lai, C.W. Chen, Discovering influencers for marketing in the blogosphere, Information Sciences 181 (23) (2011) 5143–5157.
[46] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, Proc. of the National Academy of Sciences of the United States of America (PNAS) 102 (33) (2005) 11623–11628.
[47] M. Mathioudakis, N. Koudas, Efficient identification of starters and followers in social media, in: Proc. of the International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09), ACM, Saint Petersburg, Russian Federation, 2009, pp. 708–719.
[48] N. Memon, R. Alhajj (Eds.), From Sociology to Computing in Social Networks – Theory, Foundations and Applications, vol. 1, Lecture Notes in Social Networks, Springer, 2010.
[49] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: Proc. of the ACM SIGCOMM International Conference on Internet Measurement (IMC'07), ACM, San Diego, CA, USA, 2007, pp. 29–42.
[50] R. Monclar, A. Tecla, J. Oliveira, J.M. de Souza, MEK: using spatial–temporal information to improve social networks and knowledge dissemination, Information Sciences 179 (15) (2009) 2524–2537.
[51] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, J. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science 328 (5980) (2010) 876–878.
[52] K. Musiał, K. Juszczyszyn, Properties of bridge nodes in social networks, in: Computational Collective Intelligence, Semantic Web, Social Networks and Multiagent Systems, vol. 5796, Lecture Notes in Computer Science, Springer, 2009, pp. 357–364.
[53] J.P. Onnela, F. Reed-Tsochas, Spontaneous emergence of social influence in online systems, Proceedings of the National Academy of Sciences 107 (43) (2010) 18375.
[54] A. Perer, B. Shneiderman, Balancing systematic and flexible exploration of social networks, IEEE Transactions on Visualization and Computer Graphics 12 (5) (2006) 693–700.
[55] D. Rafiei, S. Curial, Effectively visualizing large networks through sampling, in: IEEE Visualization Conference 2005 (VIS'05), IEEE, Minneapolis, MN, USA, 2005, p. 48.
[56] D.M. Romero, W. Galuba, S. Asur, B.A. Huberman, Influence and passivity in social media, in: Proc. of the International Conference Companion on World Wide Web (WWW'11), ACM, Hyderabad, India, 2011, pp. 113–114.
[57] T.A.B. Snijders, G.G. Van de Bunt, C.E.G. Steglich, Introduction to stochastic actor-based models for network dynamics, Social Networks 32 (1) (2010) 44–60.
[58] X. Song, Y. Chi, K. Hino, B. Tseng, Identifying opinion leaders in the blogosphere, in: Proc. of the ACM International Conference on Information and Knowledge Management (CIKM'07), ACM, Lisbon, Portugal, 2007, pp. 971–974.
[59] A. Stewart, E. Diaz-Aviles, W. Nejdl, L. Balby Marinho, A. Nanopoulos, L. Schmidt-Thieme, Cross-tagging for personalized open social networking, in: Proc. of the ACM Conference on Hypertext and Hypermedia (HT'09), ACM, Torino, Italy, 2009, pp. 271–278.
[60] M. Szomszor, I. Cantador, H. Alani, Correlating user profiles from multiple folksonomies, in: Proc. of the ACM Conference on Hypertext and hypermedia (HT '08), ACM, Pittsburgh, PA, USA, 2008, pp. 33–42.
[61] J. Travers, S. Milgram, An experimental study of the small world problem, Sociometry (1969) 425–443.
[62] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, B.Y. Zhao, User interactions in social networks and their implications, in: Proc. of the ACM European Conference on Computer systems (EuroSys'09), ACM, Nuremberg, Germany, 2009, pp. 205–218.
[63] A. Wu, J.M. DiMicco, D.R. Millen, Detecting professional versus personal closeness using an enterprise social network site, in: Proc. of the International Conference on Human Factors in Computing Systems (CHI'10), ACM, Atlanta, GA, USA, 2010, pp. 1955–1964.
[64] A. Ye, J. Lang, F. Wu, Crawling online social graphs, in: Proc. of the International Asia-Pacific Web Conference (APWeb'10), IEEE, Busan, Korea, 2010, pp. 236–242.