Research on Online Social Networks: Time to Face the Real Challenges

Walter Willinger
AT&T Labs-Research
walter@research.att.com

Reza Rejaie, Mojtaba Torkjazi, Masoud Valafar
University of Oregon
{reza, moji, masoud}@cs.uoregon.edu

Mauro Maggioni
Duke University
mauro@math.duke.edu

ABSTRACT

Online Social Networks (OSNs) provide a unique opportunity for researchers to study how a combination of technological, economical, and social forces have been conspiring to provide a service that has attracted the largest user population in the history of the Internet. With more than half a billion of users and counting, OSNs have the potential to impact almost every aspect of networking, including measurement and performance modeling and analysis, network architecture and system design, and privacy and user behavior, to name just a few. However, much of the existing OSN research literature seems to have lost sight of this unique opportunity and has avoided dealing with the new challenges posed by OSNs. We argue in this position paper that it is high time for OSN researcher to exploit and face these challenges to provide a basic understanding of the OSN ecosystem as a whole. Such an understanding has to reflect the key role users play in this system and must focus on the system's dynamics, purpose and functionality when trying to illuminate the main technological, economic, and social forces at work in the current OSN revolution.

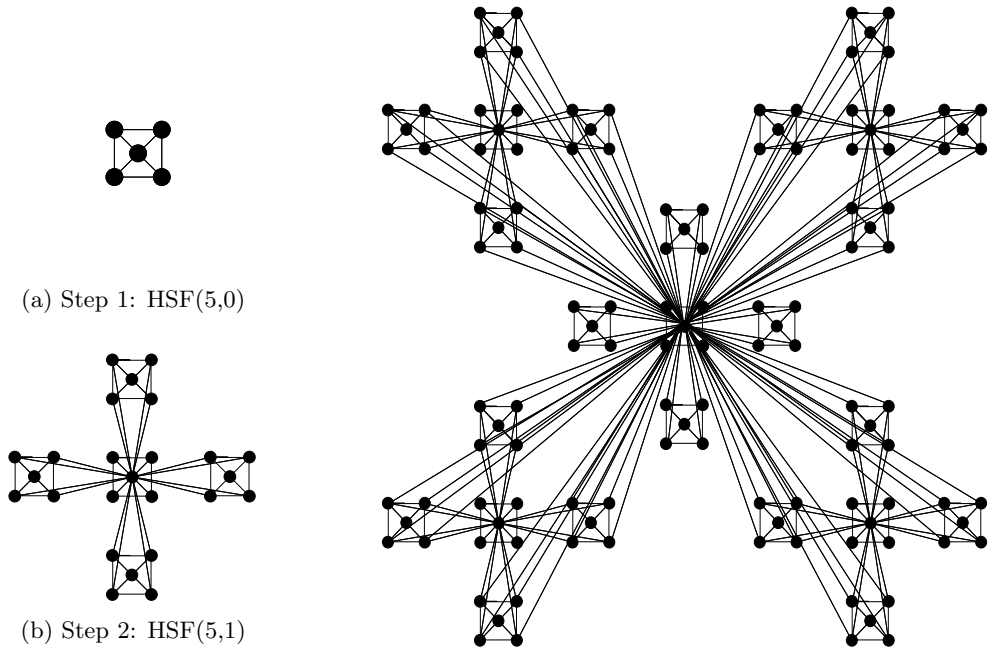
1. INTRODUCTION

To provide perspective, consider one of the more popular OSNs, Facebook. Launched in 2004 and opened up to the general public in 2006, its user base is close to an estimated 200 million as of early 2009, with a reported growth rate of close to 300,000 new users per day. Its distributed infrastructure employs CDNs and consists of about 10,000 servers worldwide (as of early 2008), with reported plans to add another 50,000 servers over the next few years [2]. As of late 2008, Facebook reportedly served over 300,000 images per second and hosted 10 billion photos in total. Since each photo is stored in 4 different sizes, this translates into some 40 billion of stored files, requiring just over 1 PB of photo storage alone [1].

Despite these staggering numbers, only a minuscule number of OSN research papers have considered issues such as Facebook's system design and architecture, user-perceived performance, active user population and network dynamics, churn and user interactions, or user behavior and privacy issues (see for example [6] and references therein). Instead, inspired by recent developments in the new field of "Network Science", the vast majority of OSN research papers has largely ignored much of the readily available domain knowledge in this area. It has focused on simple connectivity structures such as inferred friendship graphs and much of the characterization and modeling work involving these "large-scale and complex" network structures has relied on techniques from the tool box offered by "Network Science". A hallmark of these techniques is that they tend to focus on graph metrics such as node degree distribution, clustering coefficient, density, diameter, or betweenness centrality that are purely descriptive in nature (e.g., see [3, 7] and also the discussion in [10]). As such, they say little or nothing about the graphs' actual structure or dynamics. More importantly, they reduce OSNs to generic and static, and hence relatively uninteresting networked systems.

However, real-world OSNs are by nature highly dynamic structures. For example, in addition to the dynamics that is due to new users joining the system (generally by creating a new account) and existing users leaving the system (though typically without actively announcing their departure or closing their account), there is also the dynamics that results from active users interacting with each other. Here we focus on the former and note that at any point in time, the active users of an OSN may be just a fraction of all the users that have joined the network in the past. In fact, the user-based churn may be so significant as to render any friendship graph that results from inferred user relationships aggregated over long periods of time meaningless or non-informative. OSNs treat users as first-class citizen (e.g., they are the creators of content), and understanding their role, capabilities, interactions (direct or indirect, e.g., via posted photos), behavior, and dynamics is essential to study the OSNs' impact on the Internet. Moreover, users can participate in different OSNs and migrate between different OSNs, thus creating a competitive marketplace or eco-system where system design, user-perceived performance, offered services and applications, and privacy issues impact an OSN's popularity, user base, and growth rate and ultimately contribute to or determine whether the OSN can survive in this competitive environment.

To become more relevant and provide a solid understanding of the impact of the current OSN revolution on key socio-technological and socio-economic aspects of the Internet, we argue in this paper that OSN research has to change course and has to do so quickly. In particular, we maintain that future OSN research has to (1) be user-centric rather than user-agnostic, (2) abandon the traditional treatment of OSNs as static networks and become serious about dealing with the full-fledged dynamic nature of actual OSNs, and (3) give up on traditional descriptive modeling approaches that have proven to be little more than relatively uninteresting data fitting exercises. To succeed, such a revamped OSN research agenda will necessarily have to be accompa-



(a) Step 1: HSF(5,0)

(b) Step 2: HSF(5,1)

(c) Step 3: HSF(5,2)

Figure 1: 3-step construction of the (static) friendship graph F-TOYFB associated with our toy OSN TOYFB.

nied by parallel re-evaluations of currently pursued measurement techniques and data collection efforts and by simultaneous efforts aimed at developing novel and non-traditional tools for mining, analyzing, characterizing, and modeling the next-generation OSN measurements. In addition, the agenda will have to be more inter-disciplinary by actively involving domain experts from the social sciences, mathematical sciences, and engineering.

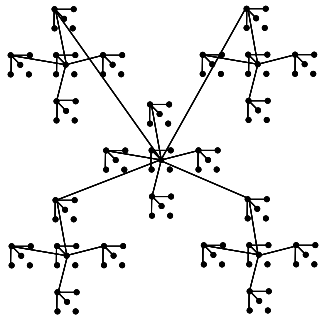
Caveat emptor: One the one hand, we do not claim that the highly stylized toy OSN considered in this paper has anything in common with real-world OSNs. We only use it to illustrate some of the main issues that arise when studying OSNs from a purely static versus a truly dynamic point of view. On the other hand, the toy OSN used in this paper has been chosen for a reason. It exhibits many of the properties that have featured prominently in studies of inferred OSN friendship graphs that have been reported in the OSN literature in the past; *e.g.,* power-law node degree distribution, local clustering, hierarchy, small-world property, or low diameter. However, only future measurements of the detailed dynamics of real-world OSNs will indicate if our toy model described below has more than just educational appeal and is indeed of practical relevance.

2. IT'S ALL ABOUT DYNAMICS!

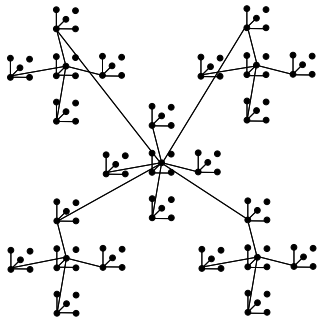
Traditional OSN research has focused mainly on inferred friendship graphs and has treated OSNs by and large as static systems. To illustrate the importance of one of the aspects of their dynamic nature (i.e., new users joining and old users leaving the system), we consider the following simple toy OSN, denoted by TOYFB. To start, we assume that the static friendship graph F-TOYFB associated with our toy OSN TOYFB belongs to the family of *hierarchical scale-free (HSF) networks* introduced and analyzed in [4]. These networks are characterized by a pair of parameters (n,m): n denotes the number of nodes in a fully meshed *cell*, and m denotes the number of levels in the hierarchy. The construction proceeds by generating n cells of size n and connecting them in a certain way. Using the thus generated structure as a new cell, the process is repeated m times to obtain an HSF(n, m) graph with m well-defined levels of hierarchy. Figure 1 shows the 3-step process of constructing an HSF$(5, 2)$ network, with 312 (bi-directional) links representing all the friendship relationships that have ever been established among the $n^{m+1} = 125$ users. Thus, like in many real-world OSNs, friendship relationships are not actively "de-activated", with the result that the friend list of any user in our toy OSN TOYFB does not decrease over time. HSF networks can be shown to have a power-law node degree distribution, rich local clustering properties (i.e., the clustering coefficient at a node with degree k follows a scaling law of the form $1/k$, meaning that the higher a node's degree, the smaller its clustering coefficient), and a well-defined "cluster-within-cluster" structure [4].

2.1 Beyond friendship graphs

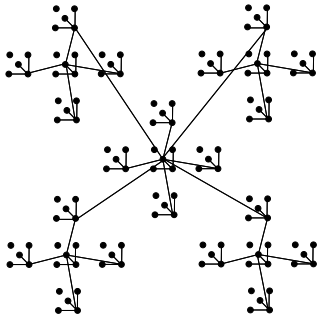
To demonstrate how static friendship graphs such as F-TOYFB can result from a very simple temporal dynamics, we describe in the following a very elementary and admittedly highly stylized evolutionary process by which the 125 users join our toy OSN, befriend other users, or become inactive by "de-activating" existing friendship relations (but without



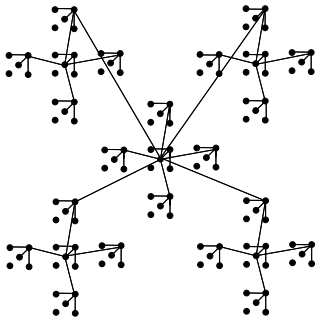
(a) t=1/16



(b) t=2/16



(c) t=3/16



(d) t=4/16

Figure 2: Specification of the dynamics of our toy OSN TOYFB **at time points** $t = 1/16$**,** $2/16$**,** $3/16$**, and** $4/16$**.**

updating one another's friend lists). They can do so at a finite number of points in time in the interval $[0, 1]$; that is, at times $t \in T = \{1/16, 2/16, \ldots, 15/16, 16/16\}$.

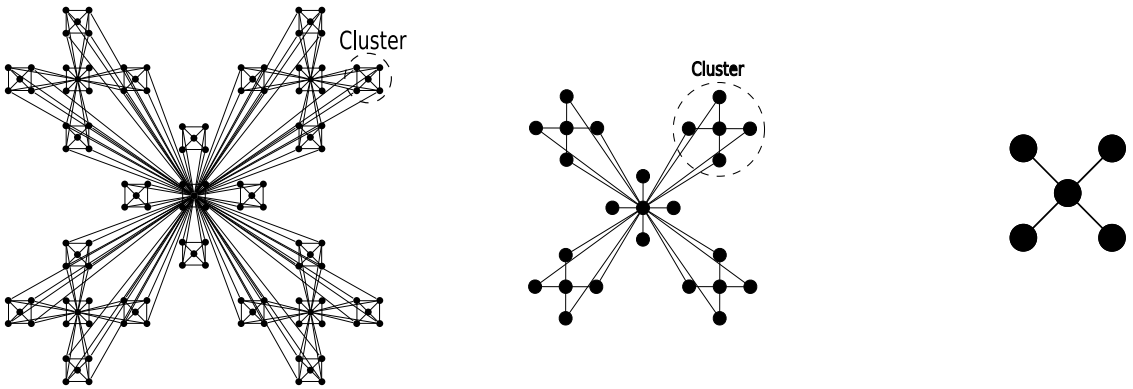
The evolution of TOYFB up to time $t = 1/4$ is illustrated in Figure 2 and proceeds as follows. At time time $t = 1/16$, the active (mutual) friendship relationships are shown in Figure 2(a). Next, at time $t = 2/16$, all the friendship relationships from time $t = 1/16$ (with one exception) are de-activated and the newly established links are shown in Figure 2(b). This process is repeated at time $t = 3/16$ (see Figure 2(c)) and $t = 1/4$ (see Figure 2(d)). The rest of the evolutionary process proceeds in three identical stages, between times $(1/4, 1/2]$, $(1/2, 3/4]$, and $(3/4, 1]$, with the only exception that the connections to the center node in the center cell of the HSF$(5, 2)$ graph are made from the left, bottom, and right generic cells, respectively.

Careful accounting of the friend lists of all the 125 users in our toy OSN shows that the static friendship graph that would result from a full crawl of TOYFB is identical to F-TOYFB, mainly because the friend lists maintained by the users do not reflect any "de-activation" of friendship relationships. Also note that at any time t, the "active" part of TOYFB, i.e., F-TOYFB(t) is (strongly) connected and that the entire toy OSN consists of this connected component and a number of singletons (i.e., sets consisting of individual users). At the same time, none of the temporal "snapshots" F-TOYFB(t) of our toy OSN exhibits any of the properties that makes their static counterpart F-TOYFB so special. Clearly, ignoring the temporal dynamic of our toy OSN as captured by the snapshots F-TOYFB(t), $t \in T$, and focusing instead on the static friendship graph F-TOYFB is bound to produce network characteristics and result in models of the OSN in question that are of little practical relevance and are likely to be misleading, if not wrong. But what are effective and efficient methods for accurately capturing and systematically characterizing the dynamic nature of large-scale real-world OSNs?

2.2 How to sample if you must

Being largely measurement-based, OSN research has relied heavily on data obtained from third-part crawlers, specifically designed to exploit OSN-specific features and services (e.g., open API) to extract user-related information (e.g., friend list) that is otherwise only available to the owner of the OSN. However, we have seen that the capture of, say, the friend lists of users, regardless of how extensive or complete, provides only very limited and inaccurate information about which users are active in the OSN at what time(s) [14]. To get a handle on the actual dynamics of a real-world OSN, any crawl has to provide some type of timing information that can be used to infer active users and distinguish them from users that were active at some point in the past but have long since moved on to other OSNs. It should be apparent that crawling our toy OSN TOYFB and extracting the timing information of link de-activations (assuming it is available and accurate) would enable us to infer the true dynamics of our toy OSN.

However, crawling a system like Facebook, with an estimated 200 million users (here, a user is identified in terms of the account she opened when joining Facebook) is not feasible. A full crawl is prohibitive due to the system's size, and any partial crawl requires proof that the obtained data is representative and not biased. The main sources



(a) Scale 1 (b) Scale 2 (c) Scale 3

Figure 3: Multi-scale representation of the friendship graph F-TOYFB associated with the toy OSN TOYFB.

for bias include spatial heterogeneity (e.g., highly skewed node degrees in the inferred friendship graph), temporal dynamics or churn (e.g., new users joining the system), and very limited timing information about user activity. Moreover, OSNs typically limit the number of third-party queries which adds yet another wrinkle to crawling OSNs for the purpose of obtaining user-specific data that goes beyond pure connectivity (i.e., friendship) information. While the majority of currently available OSN measurements results from some type of partial crawling, proof that they are representative and unbiased is unfortunately thoroughly lacking.

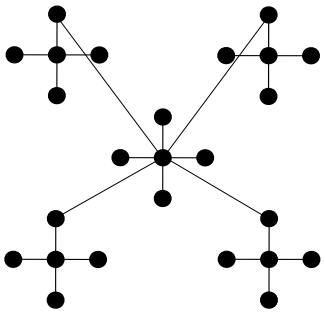
Assuming OSNs will continue to treat most user-specific information as off-limits to third-party crawling activities, OSN researchers will have to re-think OSN measurement, all the way from what to measure to how to obtain the desired measurements. The ultimate objective should be nothing less than high-quality data in support of a clear understanding and detailed characterization of the structure, dynamics, and user behavior associated with real-world OSNs. Given the problems with full or partial crawls, a promising approach to OSN measurement is *sampling* [11, 12], but even its use is far from straightforward. In theory, the goal of sampling an OSN is to obtain a "representative" sample of users and their various attributes. However, in view of an unknown underlying graph structure that is changing over time, what is meant by a "representative" sample? Even if the latter is adequately defined, obtaining such a sample inevitably requires some crawling technique, which in turn has to be informed by appropriate and up-to-date connectivity or activity information provided by the users encountered during the crawl. Given that the friendship graphs that have been used in the past in this context are inadequate (see the toy OSN example above), what is the proper (i.e., reliable and useful) information a crawler should extract from each encountered user?

2.3 Multi-scale to the rescue?

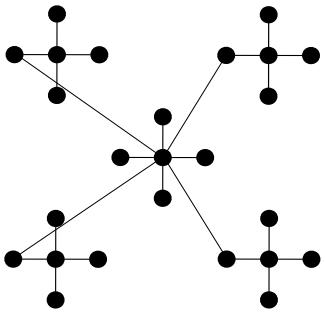
Given the sheer number of users and possible links in an OSN of the size of Facebook, the general lack of timing information needed to distinguish between active and inactive users, and the overall ambiguity associated with using available OSN measurements to infer how users interact with one another over time makes studying real-world OSNs in a rigorous and principled manner look like a near impossible task. Moreover, maybe the kind of OSN research performed to date is all that can be expected under these circumstances. Not satisfied with this status-quo, we argue here that more creative and innovative OSN research efforts should and can be pursued despite or because of the very challenges listed above.

We base our optimistic outlook on two key observations. First, snapshots taken over relatively short periods of time of inferred friendship or interaction graphs of real-world OSNs tend to exhibit pronounced clustering at different spatial scales. Tightly connected nodes in the original graph form apparent clusters at a coarser granularity, which in turn form new clusters at a yet coarser level of resolution. Moreover, graphs with different node or link attributes may cluster differently at different scales, possibly even with different overlapping clusters. Second, the temporal dynamics of a graph structure at a fine (spatial) granularity is typically faster and more noisy than the dynamics of that graph at coarser levels of resolution. Together, these two observations suggest a promising novel approach for investigating large-scale, dynamic, annotated graphs arising in the context of OSNs: Start at a coarse scale where the graph's size is small and its dynamic is slow and use the insight gained at that scale to study the graph at the next finer levels of resolution.

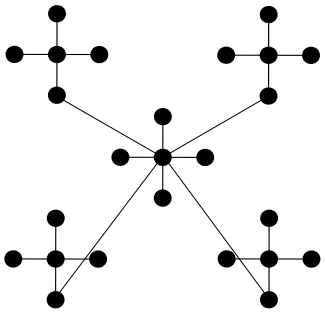
In effect, we propose here to pursue a full-fledged *Multi-Resolution Analysis (MRA)* of large-scale and evolving graph representations of real-world OSNs, and our method-of-choice for enabling and developing such an MRA is a recently proposed technique called *Diffusion Wavelets (DW)* [5, 8, 9]. Diffusion Wavelets provide a mathematical tool for performing a principled multi-scale analysis of graphs and of functions defined on graphs. They are a natural choice for constructing and studying static graphs at different (spatial) levels of resolution, for separating structure from noise, and for tracking the evolution of graphs at different (spatial and temporal) scales, where the (spatially) coarse-scaled counterparts of the original graph can be expected to evolve in time with different levels of predictability. Instead of tracking single users with all their variations in the original graph, the proposed MRA method enables the systematic tracking



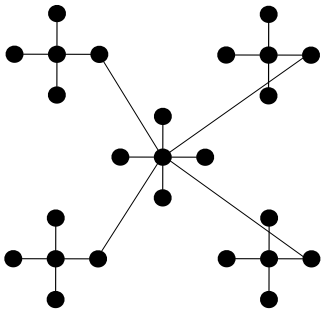
(a) t=1/16



(b) t=5/16



(c) t=9/16



(d) t=13/16

Figure 4: Slow temporal dynamics of the F-TOYFB graph at scale 2 at time points $t = 1/16$, $5/16$, $9/16$, and $13/16$.

of scale-dependent "soft" clusters identified and constructed with the help of DW. In short, the DW technique promises to support an original and practically useful MRA for large-scale evolving graph structures that reveals the critical, but possibly different structural features and forces at work at the different scales in space. Moreover, it enables a systematic study of the dynamics of the underlying structure or their coarse-scale counterparts at different scales in time.

To illustrate the main ideas, consider again the dynamic toy OSN TOYFB and its static counterpart F-TOYFB and note that one of the educational appeal of the HSF graphs is that applying to them the type of MRA proposed above can proceed by visual inspection, without the need to explain the mathematics behind DW. In particular, when viewed at the finest scale (i.e., scale 1), the HSF$(5, 2)$ graph (see Figure 3(a)) reveals an obvious cluster-within-cluster structure that suggests a natural multi-scale representation of the HSF$(5, 2)$ graph at coarser scales 2 and 3 as given by the graphs in Figure 3(b) and Figure 3(c), respectively (note that the coarsest scale 4 can be thought of as consisting of a single node but is not shown). A key feature of this successive coarsening of the HSF$(5, 2)$ graph is that every node at a coarse scale can be uniquely identified with a subset of nodes and edges at a finer scale, resulting in a natural MRA representation of HSF$(5, 2)$ that captures its visually striking hierarchical structure. The main motivation for relying on the DW technique as our method-of-choice for a MRA of graphs is that it provides the necessary mathematical framework for performing the above intuitive graph coarsening process in a rigorous manner for general graph structures.

The potential of using an MRA of graphs for studying dynamic graphs can be seen when comparing Figures 2 and 3. First, at the coarsest (spatial) scale (i.e., scale 3, Figure 3(c)), the dynamics is indeed the slowest; in fact, at scale 3, the graph structure as shown in Figure 3(c) is fixed and does not evolve over time. Moving to the next finer spatial scale (i.e., scale 2, Figure 3(b)), we observe those effects of the dynamic nature of TOYFB that manifest themselves in how the generic peripheral HSF$(5, 0)$ cells as a whole (and within their respective HSF$(5, 1)$ cells) connect to the center node of the HSF$(5, 2)$ graph; how individual users within those cells connect to one another or to the center node of their HSF$(5, 1)$ cell remains invisible at this scale. As shown in Figure 4, this particular dynamics is no longer constant, but can be fully described by specifying it at the coarse temporal scale consisting of times $t \in \{1/16, 5/16, 9/16, 13/16\}$. Lastly, to recover the full dynamics of TOYFB down to the level of how the individual users within each generic HSF$(5, 0)$ cell establish connectivity within their cell or with their center of their HSF$(5, 1)$ cell, we have to bring in the finest spatial scale (i.e., scale 1) as well as the finest temporal scale (i.e., $t \in T$).

Despite being at best a simple caricature of how real-world OSNs evolve in time, our toy OSN example has sufficiently rich structure and dynamics to illuminate how an approach to understanding evolving graphs based on a purposeful multi-scale decomposition may work for more general dynamic graphs. One attractive feature of such a decomposition is that it associates the different components of the overall dynamics with the different graph representations at the different spatial and temporal scales. Another appealing property that is highlighted by this example is that the approach can tolerate different amounts of noise or

measurement errors in precisely how the nodes at the different spatial scales establish links between them – the coarser the scale, the larger the noise or measurement errors that can be tolerated.

3. CONCLUSION AND OUTLOOK

Past research involving measurement, analysis, characterization, and modeling of OSNs has largely ignored the fact that OSNs are highly dynamic systems. To rectify this failure, we argue in this position paper that future OSN research will have to do away with much of what are presently considered to be standard and commonly-accepted measurement, modeling, analysis, and validation approaches. It will have to replace them by new methods that can account for the full-fledged dynamic features exhibited by real-world OSNs and we have listed some initial attempts. For example, while there has been some initial progress using sampling to measure heterogeneous and dynamic graph structures [11, 12], the largely unknown nature of the dynamics and churn of real-world OSNs remains a serious obstacle towards extracting truly informative measurements from today's OSNs. However, without overcoming this problem, future measurement-driven OSN studies will continue to lack the solid foundation necessary for providing an in-depth understanding of OSNs that is grounded in high-quality measurements. In terms of their analysis, we have illustrated with a simple toy example why a mathematically sound MRA methodology for large-scale dynamic graph structures would represent significant progress towards a principled treatment of real-world OSNs (including the detection and identification of different communities-of-interest or the prediction of link or node attributes in missing data scenarios).

In addition to the intra-OSN dynamics (generated, for example, by users joining and leaving one and the same OSN), there also exists empirical evidence of users migrating from one OSN to another in significant numbers [13]. What are the features of OSNs that attract masses of new users and cause the simultaneous demise of other OSNs? The present-day OSN eco-system provides a unique opportunity to observe, study, and analyze the rise and fall of real-world OSNs and try to identify the main forces responsible for the observed inter-OSN churn. However, the current OSN research pays hardly any attention to this phenomenon, even though it is ultimately the most important in practice – what are the trademarks of OSNs that attract most users? Is it their architecture or system design, their ability to provide superb end-user performance, their innovative spirit that ensures the support and actively fosters the development of new applications and services, their genuine concern for and cutting-edge approaches to issues of user privacy and content protection, etc.? In contrast to current OSN research that has no answers to these questions, future OSN research will ultimately be judged by its very (in)ability to provide correct answers to these and related questions.

4. REFERENCES

[1] `www.facebook.com/note.php?note_id=30695603919`, Oct. 2008.

[2] `www.venturebeat.com/2008/05/09/facebook-borrows-100m-to-build-out-its-infrastructure/`, May 2008.

[3] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. *Proc. WWW'07*, 2007.

[4] A.-L. Barabasi, Z. Dezso, E. Ravasz, S.-H. Yook, and Z. Oltavi. Scale-free and hierarchical structures in complex networks. *Seventh Granada Lectures*, Spain, 2002.

[5] R. R. Coifman and M. Maggioni. Diffusion wavelets. *Appl. Comp. Harm. Anal.*, Vol. 21, pp. 53–94, 2006.

[6] B. Krishnamurthy. A measure of online social networks. *Proc. of COMSNETS'09*, 2009.

[7] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. *KDD*, Chicago, IL, 2005.

[8] M. Maggioni, J. C. Bremer Jr., R. R. Coifman, and A. D. Szlam. Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs. *Proc. SPIE Wavelet XI*, Vol. 5914, 2004.

[9] M. Maggioni, A. D. Szlam, R. R. Coifman, and J. C. Bremer Jr. Diffusion-driven multiscale analysis on manifolds and graphs: Top-down and bottom-up constructions. *Proc. SPIE Wavelet XI*, Vol. 5914, 2004.

[10] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr social network. *Proc. ACM/SIGCOMM Workshop on Online Social Networks (WOSN'08)*, 2008.

[11] A. Rasti, M. Torkjazi, R. Rejaie, N, Duffield, W. Willinger, and D. Stutzbach. Graph sampling techniques for studying unstructured overlays. *Proc. IEEE INFOCOM'09 Mini-Conference*, 2009.

[12] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Trans. on Networking*, to appear, 2009.

[13] M. Torkjazi and R. Rejaie and W. Willinger. Hot today, gone tomorrow: On the migration of MySpace users. *Proc. ACM/SIGCOMM Workshop on Online Social Networks (WOSN'09)*, 2009.

[14] M. Valafar and R. Rejaie and W. Willinger. Beyond friendship graphs: A study of user interactions in Flickr *Proc. ACM/SIGCOMM Workshop on Online Social Networks (WOSN'09)*, 2009.