Computer Communications 73 (2016) 167–175

Contents lists available at ScienceDirect

Computer Communications

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

Uncovering the fingerprint of online social networks using a network motif based approach

Alexandru Topirceanu*, Alexandra Duma, Mihai Udrescu

Department of Computers and Information Technology, Politehnica University Timisoara

ARTICLE INFO

Article history:
Available online 10 July 2015

Keywords:
Online social networks
Complex network topologies
Network motifs
Classification
Similarity

ABSTRACT

Complex networks facilitate the understanding of natural and man-made processes and are classified based on the concepts they model: biological, technological, social or semantic. The relevant subgraphs in these networks, called network motifs, are demonstrated to show core aspects of network functionality and can be used to analyze complex networks based on their topological fingerprint. We propose a novel approach of classifying social networks based on their topological aspects using motifs. As such, we define the classifiers for regular, random, small-world and scale-free topologies, and then apply this classification on empirical networks. We then show how our study brings a new perspective on differentiating between online social networks like Facebook, Twitter and Google Plus based on the distribution of network motifs over the fundamental topology classes. Characteristic patterns of motifs are obtained for each of the analyzed online networks and are used to better explain the functional properties behind how people interact online and to define classifiers capable of mapping any online network to a set of topological-communicational properties.

1. Introduction

Complex networks cover an active area of scientific research inspired largely by the empirical study of real-world networks such as communication networks, economical networks and social networks. They are classified into four major types, based on the context which they model: biological networks (e.g., metabolic networks, transcription regulatory networks, protein–protein interaction networks, protein structure networks, neural networks, ecological networks, and natural food chains) [1,15,50], social networks (e.g. friendship networks, citation networks, voter networks, world markets, and political structures) [36,42,50], technological networks (e.g., computer networks, electrical circuits, and road networks) [1], and semantic networks (e.g. word-net [31] and recipe networks [43]). Without exception, all these networks can be represented as graphs, which include a wide variety of subgraphs. One fundamental property of networks are the so-called network motifs, which were introduced by Milo et al. [33]. They represent recurrent and statistically significant subgraphs or patterns in these complex networks. The fact that motifs repeat themselves in specific networks, or even among various networks, is highly correlated with the concepts of evolutionary theory. Each of these subgraphs, defined by a particular pattern of interactions between graph nodes, may reflect a framework in which particular functions are achieved efficiently. Motifs are considered to have a notable importance today because they may reflect underlying functional properties [30]. In light of their ability to uncover structural design principles of complex networks, motifs have been slowly adopted from Systems Biology into the broader perspective of Network Science. Although they foster a deep insight into the functional abilities of a network, their detection is computationally challenging even by current standards.

Particular research has been done in the areas of biology and genetics where motifs are associated with functional roles of transcription regulation networks which control the expression of genes [2]. Experimental studies show how motifs serve as basic building blocks of transcription networks. Another example is the understanding of how some cellular components are conserved across species but others evolve rapidly [54]. A notable study brings forward this new motif-inspired paradigm to uncover drug development strategies that help in the identification of drug target candidates [12]. A similar scientific track to our proposal is presented by Wang et al. in a study focused on detecting important nodes, not through the classic centrality metrics approach, but through specific motif patterns[49].

While conceptually (and functionally), complex networks can represent biological, technological, social or conceptual relationships between entities, we propose a motif-based analysis of networks from the topological perspective. As such, the fundamental topological families are: regular networks, random networks, small-world networks and scale-free networks [50]. Regular [8] and random

* Corresponding author. at Faculty of Automation and Computers, Politehnica University of Timisoara, Bd. Vasile Parvan 2, 300223, Timisoara, Romania. Tel.: +40 256 123456.
E-mail address: alext@cs.upt.ro (A. Topirceanu).

http://dx.doi.org/10.1016/j.comcom.2015.07.002
0140-3664/

networks [16] represent the basics of complex networks. The effort to mathematically express accurate and realistic models of natural phenomena (e.g. social influence, collaboration, and internet communication) has been triggered by the observation of the three fundamental properties of complex networks: average path length, clustering coefficient and degree distribution [42,50]. The well-known models of small-world [51] and scale-free [4] networks both present these network properties. Since their introduction to the literature, a considerable amount of new networks have been added, yet all fall into one of the two categories: small-world or scale-free. To recreate natural processes with a higher fidelity, there are proposals which add the small-world property to scale-free models [18,20,28], or ones that add power-law degree distribution to the small-worlds [9,22,48,55].

Our work stems from an initiative to bring the concept of network motifs closer to the field of social networks analysis (SNA) and define a new way at looking at social graphs [14]. We bring substantial new insight in terms of the types of motifs analyzed, the size and number of real-world datasets and the results and conclusion based on this new research. Thus the motivation of this paper is to provide an analytical perspective over existing state of the art complex topologies using an novel approach – classification using the network structure, namely through network motifs.

In the second part of this paper, we apply this novel perspective to differentiate between online social networks. We use empirical data to demonstrate how real social networks can be classified with different levels of appurtenance to the four topological models. Even though similar in nature, it is shown in this paper that Facebook networks, Twitter networks and Google Plus networks have very distinct topological features, as revealed by the motif-based analysis. This points out to the different features the three social platforms have in the real world.

We set out to measure the motif distributions of sizes 3 and 4 on a comprehensive database of undirected online social networks. For this, we obtain encouraging results regarding the particular patterns each of the three mentioned online platforms reveals. Their fingerprint is highly visible in terms of distribution of triadic closures, which is correlated with the clustering of nodes and short paths in the graph. The mark of triads is important as it has been shown to drive the scaling and emergence of social networks in general [24]. Also, using our approach to reveal triadic closure formation is correlated with the predictability of evolving contacts in human proximity networks [40], an important aspect of modern communication frameworks. The classifiers we obtain for each of the three online social network classes are mapped onto the four topological families and also provide a new methodology of identifying key functional properties for new network data.

1.1. Motivation and outline

In light of the general concept-driven approach to complex and social networks analysis, we propose a new perspective of looking at networks from their topological point of view. This perspective is conceptualized in Fig. 1 using the four main complex network classes: regular, random, small-world and scale-free and is provided by in-depth network motif analysis. Thus, we bring forth the following main contributions:

- Large-scale computational generation and motif distribution analysis for the synthetic topology classes. We obtain a distinct motif pattern for each such class.
- Comprehensive motif analysis of online social networks (Facebook, Twitter, and Google Plus) from which we obtain three quantifiable characteristic motif fingerprints.
- Mapping and similarity assessment of empirical networks onto topology classes, and defining a general methodology for such an approach.

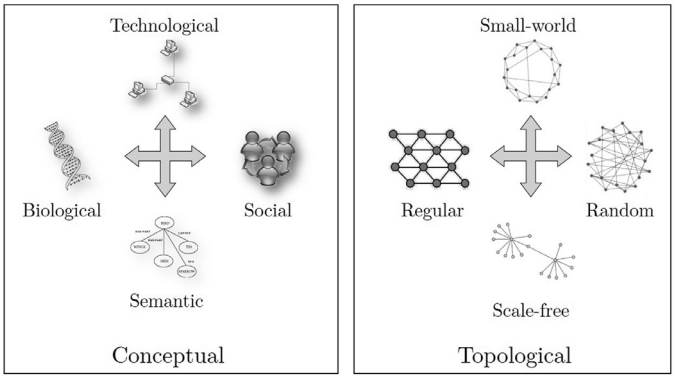

Fig. 1. The two classifications of complex networks: the conceptual perspective versus the topological perspective.

- Correlation and discussion of the individual motifs that occur in each fingerprint, and an outlining of the functional properties behind the three online social platforms.

2. A new perspective over the related work

Comparing complex networks is aimed at a deeper understanding of the interaction patterns between these systems [4,42,51], and extracting their common properties helps improve the models even further [3,23,51]. However, the predominant method of graph metric comparison suffers from limited information [27]. Some notable means of comparison are the distance ratio measure [7], used to compare individual mental models, a comparison from the data analysis perspective [27] and the study of the self-similarity of complex networks [41]. The network dimension is a key feature in understanding not only network topology, but also dynamical processes on networks, such as diffusion, percolation and other critical phenomena [13]. The fractal dimension d_B is proposed based on the belief that social networks are not invariant or self-similar under a length-scale transformation. Fractal dimension has been measured on multiple varied real world networks like the WWW, biological networks, and actor networks, and we will use it as an alternative to the standard metric comparison.

From a topological perspective there are studies done both in the direction of classifying social network models [23] and of structural pattern detection [37]. These methods however serve a higher level of meta-analysis rather than as measures of similarity.

The work done in the field of network motifs, since their introduction [33], has seen the definition of several super-families of evolved and designed networks by the same authors [32]. They present families of complex networks grouped together by the similar significance profiles (SP) of motifs in the networks compared to the normal occurrence in random networks. These families include:

- direct transcription interactions (in bacteria and yeast);
- signal-transduction interactions (cell signaling, neural networks);
- web hyperlinks and social networks;
- word-adjacency networks networks (in English, Spanish, Japanese).

Another study shows an alternative approach to the analysis of community structure by partitioning a network into a *core* of high degree nodes that are highly interconnected to each other, and a *periphery* of nodes that are not so well connected. The core has an important role in mediating most of the minimum path length motifs and has an integrative aspect over the topology [35].

With great preponderance, all studies revolve around the classification of networks – empirical or synthetic – from the conceptual point of view, into one of the mentioned four main categories.

However, many of the functional properties inherent to the different classes of complex networks stem from their underlying topological features.

In accordance to our previous work [14,44], we propose an alternative perspective, in which we consider any emergent complex network a mixture of the four fundamental topology classes: random, regular, small-world and scale-free. In the context of social networks, for example, it is a well know fact in the literature that characteristics of each topology are present. Collaborations, sexual interactions, friendships, and citation networks are good examples of scale-free networks [34,50]; voter networks, influence networks, food chains, and human communities are examples of small-worlds [15,50], and they feature properties of regular organization and/or random long range links as well. Our main motivation for reclassifying networks based on topology is driven by the fact that each network model can be characterized by a certain mixture of topological properties. We find out that this mixture of properties creates specific patterns over which apparently diverse networks can overlap. By applying this methodology on online social networks we bring an original contribution of how we can do social networks analysis.

The core analytical instrument with which we define the classification based on topology classes is network motifs. More specifically, given a distribution of motifs D_N over a network N one *can* classify the network into one super-family which encompasses a particular concept (e.g. social and technological), but one *cannot* associate the distribution D_N with the fundamental complex network topologies. Fig. 1 depicts the two types of classifications for complex networks. The solution to this main outline is discussed in the next section.

3. Methodology

We propose a two step approach into classifying online social networks. First, we measure the distributions of motifs of sizes 3 (i.e. subgraphs with 3 nodes) and 4 on synthetically generated networks. We have implemented the algorithms for generating regular mesh networks, Erdős-Rényi random networks [16], Watts-Strogatz small-world networks [51], and Barabási-Albert scale-free networks [4] in Gephi [5]. Gephi is a world-leading open-source large data visualization tool built on the Netbeans framework using Java. After generating a relevant amount of such networks, ranging from 100 to 5000 nodes, with parameter values characteristic for each class, we use FANMOD to run the motif detection [53]. FANMOD is a light-weight tool for fast motif detection designed using one of the fastest detection algorithms available, RAND-ESU [52]. As depicted in Fig. 2, the first step is to find the distributions $D_{reg}, D_{rnd}, D_{sw}, D_{sf}$ for the four corresponding topology classes.

All the generated and used networks are undirected and unweighted, since edges model mutual social ties with no additional information regarding tie strength, reciprocity, etc. Many studies (from the originating fields of Medicine) rely on the analysis of motifs of size 3 in directed contexts. The upper size limit is commonly imposed due to the computational complexity of detecting larger motif structures. However, since we deal with an undirected context, the processing time is greatly reduced. For example, there are 13 different combinations of motifs of size 3 in a digraph, as depicted in Fig. 3, but only 2, respectively 6 undirected motifs of sizes 3 and 4. The codes of each motif depicted in Fig. 3 are standardized in the literature and represent the serialized binary value of the adjacency matrix (row by row) converted to a decimal value. For example, code 14 originates from the matrix 000 001 110 converted to base 10. In this paper we measure the distributions of motifs depicted in Figs. 3b and 3c, and will refer to them using the corresponding codes.

The second step is to run the same process of detecting motifs and determining the distributions on the three chosen online social networks. We have chosen Facebook, Twitter and Google Plus as they are the most popular sites in this field [17,29]. The empirical data is gathered from the Stanford large network dataset collection [25] and from a comprehensive private repository populated with Facebook friendship graphs of students aged 19-25. The averaged results of running FANMOD on these networks yields the characteristic distributions D_{FB}, D_{TW}, D_{GP}.

To correlate the distribution vectors of the empirical datasets with each vector of the reference distributions we use the existing fidelity metric φ [45]. The metric is tailored to express of similarity between any two generic vectors, in a weighted or unweighted context. In this paper we use the unweighted arithmetic fidelity metric:

$$\varphi^j = \begin{cases} \frac{1}{n}\sum_i \frac{\overline{m_i}}{2\overline{m_i}-\overline{m_i^j}} & if\ \overline{m_i^j} < \overline{m_i} \\ \frac{1}{n}\sum_i \frac{\overline{m_i}}{\overline{m_i^j}} & if\ \overline{m_i^j} \geq \overline{m_i} \end{cases} \quad (1)$$

where j is the index of empirical distribution model being compared to the reference, $i = \{1, 2, \ldots n\}$ is the index of the motif which describes the two models being compared, and n is the total number of common motifs. The closer the φ metric is to 1 the more similar the models are. The measurements on the reference model are m_i, respectively m_i^j on the model being compared.

By measuring all similarities one can express each empirical distribution using one or more distributions of the four topological classes as:

$$D_j = \alpha_j^{reg} \times D_{reg} + \alpha_j^{rnd} \times D_{rnd} + \alpha_j^{sw} \times D_{sw} + \alpha_j^{sf} \times D_{sf} \quad (2)$$

where j is the index of any of the three social network distributions (i.e. FB, TW, GP; e.g. $j = FB \rightarrow D_{FB}$, α_{FB}^{reg}, α_{FB}^{rnd}, α_{FB}^{sw}, α_{FB}^{sf} etc.), or any empirical complex network in general. The coefficients α are obtained from the normalized similarities with each topology respective class. For example, α_{FB}^{reg} is the normalized similarity of the Facebook motif distribution (vector) towards the distribution found in regular networks.

In contrast to previous work [14], the motif sizes used in this study are fixed to 3 and 4, in an undirected context. While there are approaches in the literature studying network functionality using motifs of larger sizes (up to 6), we rely only on the size 3 and 4 motifs since there are few such distinct patterns, and are much more numerous to be found in graphs, and thus substantially more relevant [2].

4. Dataset analysis

The presented motif-driven methodology requires the synthetic generation of networks pertaining to each of the four topology classes (within the characteristic parameter values), and the acquisition of friendship networks for each of the three online platforms. In this section we briefly present the parameters and settings used for generating the data, as well as the graph metrics obtained for each network class.

Even though friendship graphs vary in size significantly, from as few as 100 nodes to as many as 5000 nodes, it is a known statistic that the predominant majority of such networks revolve around the size of 300 nodes [19]. We thus generate data accordingly and concentrate on synthetic networks within that range. Moreover, we rely on public data gathered from the Stanford large network dataset collection [25] which offers networks of hundreds up to millions of nodes. Taking into consideration the fact that graph size significantly impacts motif distributions, in order to enable a comparison at the same scale, all chosen synthetic networks are within the range of real-world ego-networks. The following datasets are used in this paper:

- Regular: we have generated standard 2D mesh networks of sizes 200, 300 and 500.
- Random: we have generated random networks of the same sizes using the Erdős-Rényi algorithm [16], and the wiring probabilities $p_1 = 0.05$ and $p_2 = 0.1$.

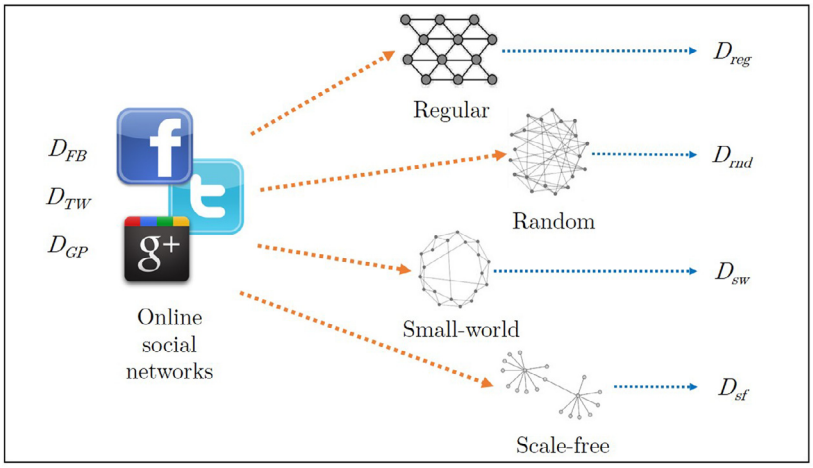

Fig. 2. The process of classifying the three online social networks (Facebook, Twitter, and Google Plus) using the four topological classes. Each motif distribution of the social networks (D_{FB}, D_{TW}, D_{GP}) is expressed as a combination of the four theoretical distributions ($D_{reg}, D_{rnd}, D_{sw}, D_{sf}$).

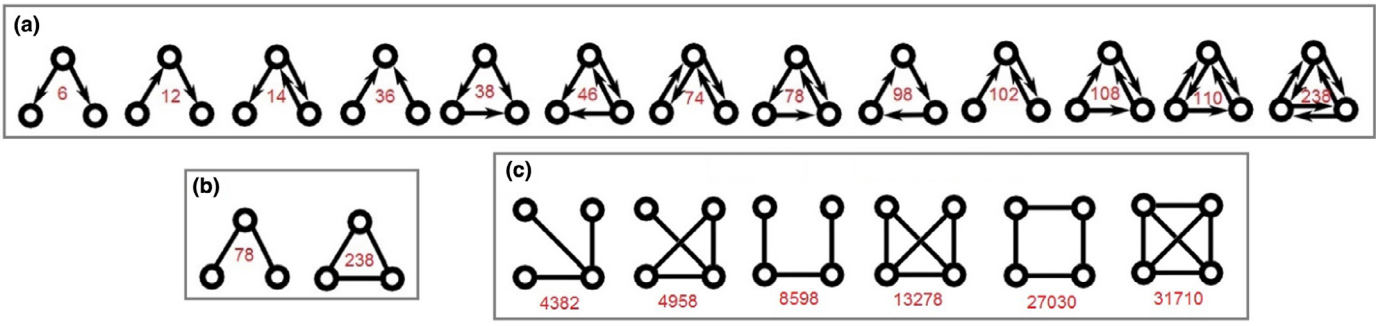

Fig. 3. Motifs representation. a. All existing motifs of size 3 in a directed graph. b. The two types of motifs of size 3 in an undirected graph. c. All existing motifs of size 4 in an undirected graph. The code of each motif corresponds to the decimal value of its serialized adjacency matrix.

- Small-world: multiple networks have been generated using the Watts-Strogatz algorithm [51], with sizes 300 and 500 nodes, wiring distance $k_1 = 2$ and $k_2 = 5$, and rewiring probability $p_1 = 0.05$ and $p_2 = 0.1$.
- Scale-free: multiple networks have been generated using the Barabási–Albert preferential attachment algorithm [4], with sizes 200, 300 and 500 nodes.
- Facebook: over 50 different friendship ego-networks have been used for metric measurements and motif analysis. Ten ego-networks are obtained from the Stanford large network dataset collection [25,26] and have a total of 4039 nodes and 88234 edges, when combined. Furthermore, we also rely on personally gathered data using the *netvizz* Facebook application [39] with which we have obtained 50 ego-networks of sizes 150–5000 nodes.
- Twitter: using the same online repository [25], 973 Twitter circles are provided. The combined network consists of 81306 nodes and 1.7M edges. For this study, we rely on 50 chosen ego-networks, with sizes within the mentioned ranges of 200–500 nodes.
- Google Plus: we use 50 ego-networks from the same study of Leskovec et al. [26]. The combined friendship network consists of 107614 nodes and 13.7M edges. The chosen networks are all within 200–500 nodes.

Measuring the representative graph metrics over the acquired data gives conclusive results for average degree (*AD*), average path length (*L*), average clustering coefficient (*C*), modularity (*Mod*), network diameter (*Dmt*), and network density (*Dns*). Table 1 shows the distribution of averaged topological properties on each network class.

Table 1
Specific values for average degree (*AD*), average path length (*L*), average clustering coefficient (*C*), modularity (*Mod*), diameter (*Dmt*), and density (*Dns*) averaged for each data set.

	AD	*L*	*C*	*Mod*	*Dmt*	*Dns*
Regular	6.63	3.34	0.065	0.05	8	0.013
Random	7.55	2.40	0.049	0.27	4	0.050
Small-world	3.99	5.61	0.321	0.73	11	0.005
Scale-free	3.12	4.60	0.015	0.62	10	0.003
Facebook	19.82	2.48	0.266	0.47	8.5	0.050
Twitter	12.39	2.68	0.239	0.28	7	0.054
Google Plus	12.15	3.90	0.404	0.44	12	0.035

5. Results and discussion

Following the methodology description in Section 2, the first result is the motif distribution on the four topology classes. The distributions D_{reg}, D_{rnd}, D_{sw} and D_{sf} are depicted in Fig. 4 and, numerically, in Table 2. Important to note is that, for each class of networks in part, we have obtained the same motif distributions regardless of network size or other specific parameters (presented in Section 3). For example, all small-worlds exhibit the same distribution D_{sw} independent of the generated network size (100–5000 nodes) and of the rewiring probability p (0.05-0.1).

By applying the same methodology on the empirical data, we obtain the distributions D_{FB}, D_{GP} and D_{TW}. These are depicted in Fig. 5 and also show very distinct fingerprints.

If we were to analyze the presented datasets from the conceptual perspective of social networks, there would be little to differentiate

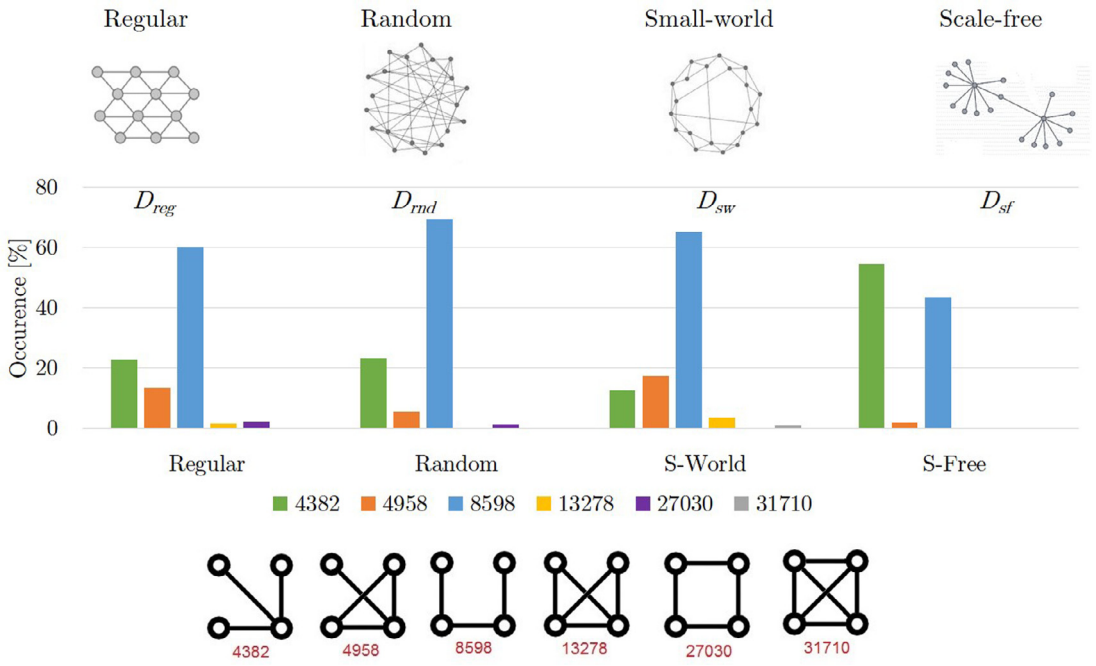

Fig. 4. The resulting motif distributions on the regular (D_{reg}), random (D_{rnd}), small-world (D_{sw}) and scale-free (D_{sf}) topologies. The occurrence of each motif is expressed in percentage in the central histogram for each network class in part. As can be seen, only distinct motifs (not all) characterize each network class. All 6 motifs of size 4 are depicted at the bottom of the figure.

Table 2
Numerical values for the distributions of the four topology classes (rows 1–4) and of the three online social networks (rows 5–7), expressed in percentages as to how often the respective size-4 motifs occur relative to the total number of recurring motifs. Each column highlights in bold the highest motif occurrence for any of the four toplogy classes (1–4).

	Motif ID:		Triads [%]			No triads [%]		
			4958 *one triad*	13278 *two triads*	31710 *four triads*	4382 *star*	8598 *chain*	27030 *rectangle*
1	Regular	D_{reg}	13.45	1.54	0.084	22.63	60.16	**2.14**
2	Random	D_{rnd}	5.613	0.26	0.004	23.25	**69.46**	1.41
3	S-World	D_{sw}	**17.46**	**3.51**	**1.08**	12.62	65.12	0.19
4	S-Free	D_{sf}	1.76	0.01	0.001	**54.39**	43.65	0.017
5	F-book	D_{FB}	32.44	11.41	5.25	17.49	31.75	1.66
6	GPlus	D_{GP}	28.86	12.33	4.14	31.48	21.34	1.84
7	Twitter	D_{TW}	27.33	11.94	6.23	22.50	30.43	1.53

and conclude, since most online social networks serve a similar purpose. However, even at a first visual impression over Figs. 4 and 5 it is interesting to point out how diverse the motif-based fingerprints of all 7 network types are. To facilitate the results discussion we also provide the numerical results in Table 2

To begin with, the conclusions based on the obtained data are that each of the four topology classes has a distinct element in its motif-fingerprint. In our discussion we reference the fact whether networks favor the formation of triadic closures more, or keep triangles open. Looking at Fig. 3c, the six motifs can be divided in two categories: motifs with triads (2nd (4958), 4th (13278), and 5th (31710)) and with no triads (1st (4382), 3rd (8598), and 5th (27030)) in their structure. Triadic closures have been found to be one of the fundamental properties that give complexity and heterogeneity to social networks [6,24]. This strongly impacts the communication through each network. By condensing the data from Table 2 we present the occurrence of the two types of mentioned motifs in Table 3.

To ease the discussion based on each motif type, we keep them highlighted in *italics* and redefine them using a more intuitive keyword. To the best of our knowledge, this useful naming is a novelty introduced in this paper and was not found in any notable previous work [14,32,33].

Table 3
Percentage of total motifs of size-4 that have triadic closures versus motifs that do not have any closed triangles in their structure, measured for each network type in part. The results are obtained through the condensation of the two sections in Table 2.

	Triads [%]	No triads [%]
Regular	15.08	84.92
Random	5.88	94.12
S-World	22.06	77.94
S-Free	1.78	98.22
Facebook	49.1	50.9
Google Plus	45.33	54.67
Twitter	45.54	54.46

- Motifs: 4382 - *star*, 8598 - *chain*, 27030 - *rectangle*.
- Motifs: 4958 - *one triad*, 13278 -*two triads*, 31710 - *four triads*.

Bridging our obtained motif distributions with the study of triads, we note that regular networks have the least characteristic mark, with a preference towards *chain* and also *star* and simple *one triad* constructs. The overall homogenous mixture reveals the fact that mesh networks keep a high local clustering (*one triad*). Overall,

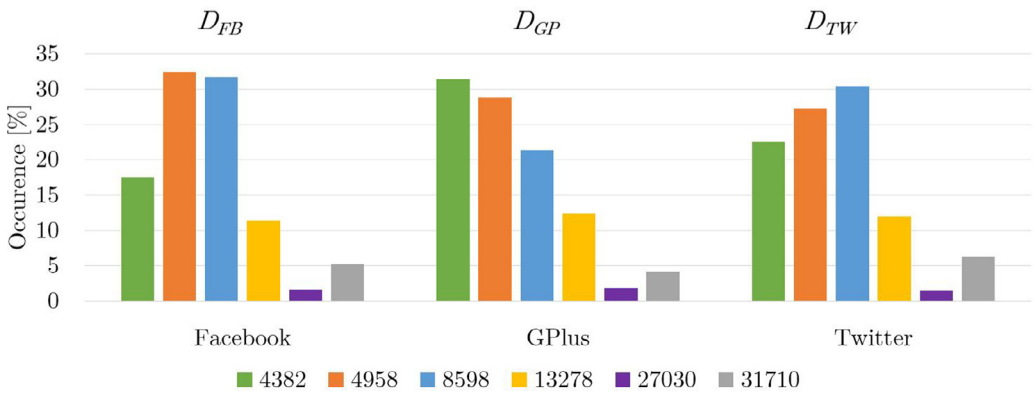

Fig. 5. The resulting motif distributions on the online social networks: Facebook (D_{FB}), Google Plus (D_{GP}), and Twitter (D_{TW}). The occurrence of each motif is expressed in percentage. As can be seen, distinct motif patterns characterize each network class. The codes of each motif are the same as the ones used in Fig. 4.

regular networks have 84.92% motifs that do not contain triads, and 15.08% motifs that contain them (from Table 3), which replicates the state of the art experiments in this field [50]. Random networks have the same high occurrence of chains, and a very specific low occurrence of *one-*, *two-* and *four triads*. Summing up the values, random networks have less than 6% triads in them, which again strengthens the known facts about low clustering in favor of a short path length. Small-worlds are a special case of empirically observed networks that lie their properties between the regular and random topologies. They favor high clustering and short path length. Our analytical approach shows a fingerprint in terms of high density of *chains*, *one triads* and especially *four triads* (over 1%). Looking also at Table 3 we notice that small-worlds are the most balanced type of topology with roughly 22% triads, and 78% no triadic formations. This balance gives them their realism in terms of replicating real social networks. Finally, scale-free networks have emerged to cover one shortcoming of small-worlds, namely the lack of preferential attachment and a power-law degree distribution, which are essential in modeling real world friendships. The scale-free network is characterized through many *chains*, but more interesting, many *stars*, and an extremely low number of *two-* and *four triads*. Added together, we can observe that there are only 1.78% motifs with triads in a scale-free network. The high occurrence of *stars* is correlated with the hub nodes with on top of the power-law degree distribution, which is specific only to this topology class.

Moving on to the empirical online social networks, we notice very distinct distributions of the six motifs (Fig. 5). Facebook friendship networks are characterized though a lower number of *stars*, but many *one triads* and *chains*. We can conclude that while there is a low tendency for hub formation (like in pure scale-free networks) the average path length is also maintained short. Complementary to previous work [14], the obtained remarks also coincide with the data presented in Table 1. Google Plus one the other hand has a relatively lower number of *chains*, and a high number of *stars* and *one triads*. This network can be interpreted as one with higher clustering and and longer path lengths. Google Plus networks are known for their community (circle) based organization. Finally, Twitter networks are the most homogenous, with many *chains*, and an average-high number of *stars* and *one triads*. This fact translates into a more regular structure due to the concept of followers, which enable the creation on many random long-range links, with a disregard towards local clustering and triadic closure formation.

Taking the analysis beyond the mere topological level, we find a correlation between the characteristic graph metric values (see Table 1) and the obtained distributions of motifs. To begin with, the prevalent occurrence of triads in the small-worlds can be explained by the higher clustering coefficient and higher modularity.

Table 4
Numerical values for the distributions of the four topology classes and of the three online social networks, expressed in percentages as to how often the respective size-3 motifs occur relative to the total number of recurring motifs.

Motif ID:		78 *chain*	228 *triangle*
Regular	D_{reg}	93.22	6.78
Random	D_{rnd}	97.37	2.63
S-World	D_{sw}	84.31	15.69
S-Free	D_{sf}	99.49	0.51
Facebook	D_{FB}	72.58	22.42
Google Plus	D_{GP}	76.87	23.13
Twitter	D_{TW}	75.28	24.72

These networks have a 15–1000 times higher concentration of*four triads* than all other topology classes. The low concentration of *stars* comes to support the lack of a power-law degree distribution. The small-world effect is mapped in the real-world network through the stronger community structure of Facebook and Google Plus networks. On the other hand, the lack of triads found in scale-free networks is a result of the power-law degree distribution. Its low clustering and relatively higher average path length are explained though the lower occurrence of *chains* and *rectangles*. The very low modularity of regular networks is correlated with the very high occurrence of *rectangles*, which suppress the formation of clear, distinguishable communities. One of the goals of social networks analysis is to create better generative models for real-world networks, thus our motif distribution – graph metric correlation may help improve the generation of specific synthetic networks. Based on these results, there are heuristic algorithms which can used to create synthetic networks with the required metric distributions. [38,46].

To enhance the visual differentiation and similarity between the obtained motif patterns we provide a radar chart overview in Fig. 6. Notable in Fig. 6a are the higher occurrence of *stars* in scale-free networks, the low preference towards triads of the scale-free and random networks. In Fig. 6b we notice a good overlap between Facebook and Twitter networks, with high occurrences of *chains* and *one triads*, while Google Plus favors more *star* formations.

In order to further validate the insightful perspective revealed by motifs of size 4, we reapply the same methodology using motifs of size 3. In support of our claims, we briefly mention that the are only two types of motifs of size 3 in an undirected context. These can be seen in Fig. 3b and we will refer to them as *chain* (78) and *triangle* (238). Table 4 contains the distribution data for each of the seven networks.

Even though motifs of size 3 have significantly less structural complexity compared to size 4, they do reveal and sustain out

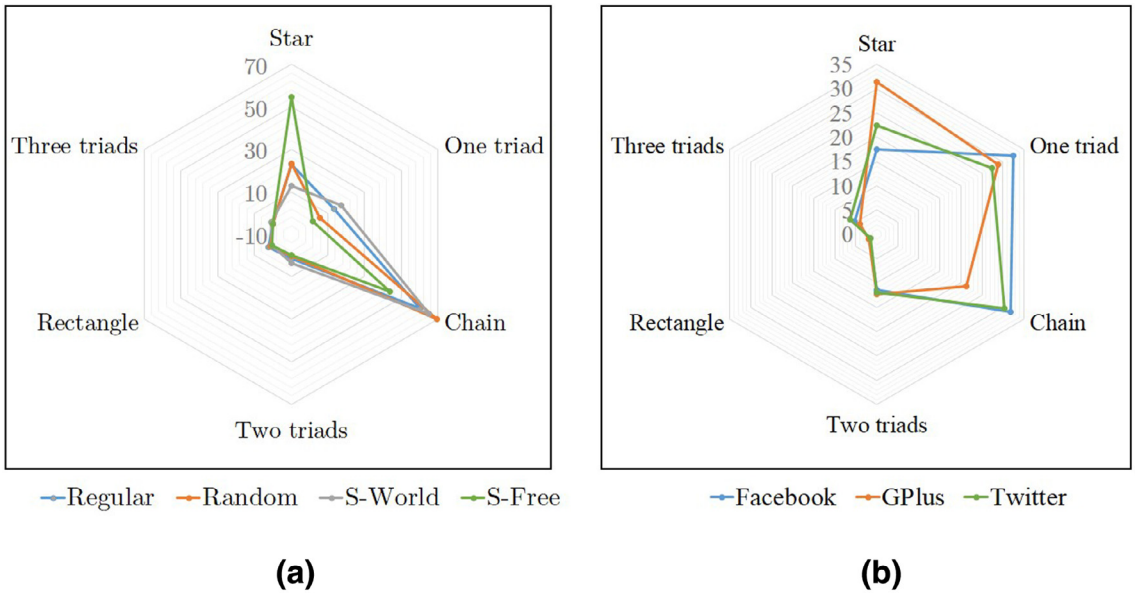

Fig. 6. Radar chart showing the 2-dimensional distribution of motifs of size 4 for the topology classes (a) and the online social networks (b).

Table 5
Similarity between the empirical network models and each topology class. The similarity is measured by applying the φ-metric on the distribution vectors as described in Eq. 1. The columns labeled n display the normalized values for the obtained similarities, according to Eq. 2. The sum of n-s is equal to 1 (100%) on each column.

	Facebook		Google Plus		Twitter	
	φ_{FB}	n	φ_{GP}	n	φ_{TW}	n
Regular	0.62	*0.266*	0.61	*0.269*	0.65	*0.267*
Random	0.60	*0.257*	0.58	*0.255*	0.65	*0.267*
Small-world	0.60	*0.257*	0.56	*0.247*	0.59	*0.243*
Scale-free	0.51	*0.219*	0.52	*0.229*	0.53	*0.219*

previous claims. Scale-free networks consistently favor open triangles to closed ones, with roughly 0.5% *triangles* in their structure. Small-worlds present the same balance between *chains* and *triangles* like in Table 3. Regular networks have a notably higher occurrence of *triangles*, and random networks of *chains*, in conformity with previous claims. Finally, Facebook, Google Plus and Twitter networks share similar distributions of *chains* and *triangles*. We note that motifs of size 3 are insufficient to assess undirected friendship graphs.

For a final overview, we apply the φ-metric on the distribution vectors of motifs of size 4 and obtain the numerical data shown in Table 5. A value of 1 means complete similarity, while a value of 0 means complete dissimilarity. The percentages of the fidelity are normalized into n-values which, summed on each column, add up to 1. The data is interpreted as, for example, Facebook can be mapped 26.6% over regular, 25.7% over random, 25.7% over small-world, and 21.9% over scale-free networks.

In interpreting the obtained fidelity results, we have to keep in mind the fact that the overall open-versus closed-triangles ratios are very similar. Specifically, this is displayed in the lower halves of Tables 3 and 4. Thus, the variations in terms of φ are small, but they map to significant structural differences [45]. Fig. 7a shows the 2-dimensional similarity mapping between the online social networks and the four topologies and Fig. 7b shows how much each topology contributes, in total, to the mapping of the three online social networks.

The fact that the highest overall occurrence is that of the regular topology, and the lowest, that of the scale-free topology, denotes an important real-world aspect of social networks: the formation of hubs is a rather exceptionally rare event, seemingly random long range links tend to form much more often, and the fundamental structure of social networks is based on mesh networks with a tendency towards local clustering. This observation sustains the fact that geographical proximity is indeed the main drive for friendships creation in society [10,47]. Furthermore, the predominantly high occurrence of *chains* in all topology classes seems to be a natural facilitator of new friendships creation. A new study proves that new friendships are preferentially created between nodes located at geodesic distances 2 and 3 in the social graph [11]. This conclusion strongly supports our results regarding *chains*which become natural pathways of length 3 between unconnected nodes. To better interpret the similarity results we corroborate the results in Table 1 with measurements of variance of the normalized fidelities (n) and conclude upon the following:

- Google Plus networks have the lowest variance (2.77-e4) showing a greater topological homogeneity. They have higher scale-free and regular appurtenances, which translates into a higher average path length (L) and a strong community structure (*Mod*). Empirically and intuitively, we explain this through the *circle* concept introduced by Google. Circles tend to offer better socializing within clusters of friends but they also limit external contacts. As most friendship clusters follow a normal distribution of contacts (degrees), the resulting model is classified as a *regular topology with preferential community formations*.
- Twitter networks have the highest variance (5.63-e4) presenting the highest topological heterogeneity. They have more notable random and regular characteristics, which translates into a very short average path length (L) and a weak community structure (*Mod*). Intuitively, we explain this through the *follower* concept specific to the Twitter online platform. The act of following tends to omit local clusters formation, or be in any way linked to geographical proximity. On the other hand, many users follow distant celebrities and/or users with same interests that are evenly spread across the globe. Uncharacteristic for real tie formations, Twitter is classified as a *heterogeneous regular topology with random long range links*.
- Facebook networks have a variance situated between the other two (4.47-e4), presenting a good mixture of all topology types. Nonetheless, they have higher small-world and regular properties, which translates into a short average path length (L) and

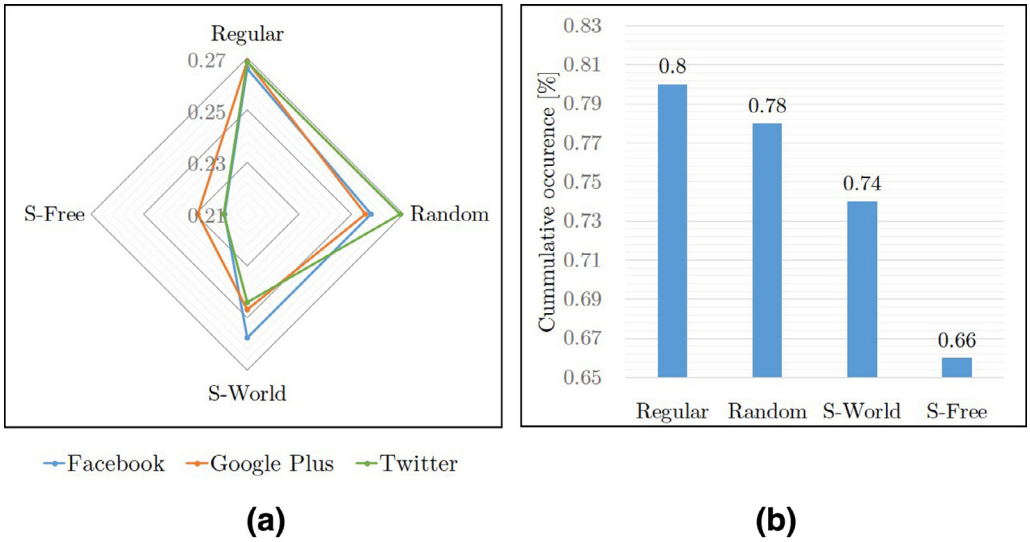

Fig. 7. a. Radar chart showing the 2-dimensional mapping of the online social networks over the four topology classes. The mapping is done using the fidelity metric φ to assess the similarities based on the distribution of size 4 motifs. b. The cumulative occurrence of each topology class obtained by adding the normalized fidelities (n) on each row (from Table 5). It shows how much each topology contributes overall to the three empirical networks.

a strong community structure (*Mod*). Based on these observations, one could say that they lie between Twitter and Google Plus. Intuitively, but also backed up by other relevant research, Facebook friendships are considered the best replica and substitute for real-world friendships [21]. This idea is further supported by the fact that their fidelity distribution also coincides with the overall fidelity distribution depicted in Fig. 7b. The stronger community structure, but with low average path lengths, seems to be a natural emerging property of the society, modeled through the *friend*-ing concept on Facebook. In fact, this seems to narrow down the distances between communities until they start overlapping. With a very characteristic real tie formation process, we classify Facebook as a *regular topology with interspersed small-worlds*.

6. Conclusions

In this paper, we have shown that studying complex networks from a topological perspective, though the insight offered by network motifs, is a new fundamental approach in understanding the emergence of social networks. Indeed, motifs highlight functional aspects of the driving forces behind online social network creation, ties formation, community emergence, and overall communication trends. Our comprehensive social networks analysis, based on graph metric and fidelity assessments, has found a predisposition for characteristics of regular networks (geo-proximity drives tie formation), followed closely by random network aspects (long range link formation), then, with diminishing predisposition, by small-world properties (tendency to cluster and close triads), and, with very low occurrence, characteristics of scale-free networks (hub formation). Finally, we have shown that each online social platform has quite distinct properties, which imply distinct motif fingerprints, and thus different communication mechanisms.

Based on our observations, and stemming from motif analysis, Facebook, Google Plus, and Twitter networks are not similar at all when it comes to mapping them over the fundamental topology classes. Each presented characteristic defines a different approach to dealing with processes like network growth, new tie formation, community formation, information diffusion and triadic closures. We believe our work will pave the way for a better understanding of the secrets that lie behind modeling and understanding dynamics in our societies.

Acknowledgments

This work was partially supported by the strategic grant POSDRU/159/1.5/S/ 137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews Modern Phys. 74 (1) (2002) 47.
[2] U. Alon, Network motifs: theory and experimental approaches, Nature Reviews Genetics 8 (6) (2007) 450–461.
[3] L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Classes of small-world networks, in: Proceedings of the National Academy of Sciences, vol. 97, 2000, pp. 11149–11152.
[4] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[5] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proceedings of ICWSM, 2009.
[6] G. Bianconi, R.K. Darst, J. Iacovacci, S. Fortunato, Triadic closure as a basic generating mechanism of the structure of complex networks, arXiv preprint arXiv (2014) 1407.1664.
[7] N. Caseiro, P. Trigo, Comparing Complex Networks: An Application to Emergency Managers' Mental Models, in: Social Simulation (BWSS), 2012 Third Brazilian Workshop on, IEEE, 2012, pp. 128–131.
[8] W.-K. Chen, Graph theory and its engineering applications, World Scientific Vol.5 (1997).
[9] Y. Chen, L. Zhang, J. Huang, The Watts–Strogatz network model developed by including degree distribution: theory and computer simulation, J. Phys. A: Math. Theoretical 40 (29) (2007) 8237.
[10] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1082–1090.
[11] F. Collet, P. Hedström, Old friends and new acquaintances: Tie formation mechanisms in an interorganizational network generated by employee mobility, Social Netw. 35 (3) (2013) 288–299.
[12] P. Csermely, T. Korcsmáros, H.J. Kiss, G. London, R. Nussinov, Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review, Pharmacol. Therapeutics 138 (3) (2013) 333–408.
[13] L. Daqing, K. Kosmidis, A. Bunde, S. Havlin, Dimension of spatially embedded networks, Nature Phys. 7 (6) (2011) 481–484.
[14] A. Duma, A. Topirceanu, A network motif based approach for classifying online social networks, in: IEEE 9th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2014 IEEE, 2014, pp. 311–315.
[15] D. Easley, J. Kleinberg, Networks, Crowds, and Markets, Cambridge Univ Press Vol. 8, 2010.
[16] P. Erd6s, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci 5 (1960) 17–61.

[17] G.A. Fowler, Facebook: One billion and counting, The Wall Street Journal 4 (2012).
[18] P. Fu, K. Liao, An evolving scale-free network with large clustering coefficient, in: 9th International Conference on Control, Automation, Robotics and Vision, 2006. ICARCV'06. IEEE, 2006, pp. 1–4.
[19] K. Hampton, L.S. Goulet, L. Rainie, K. Purcell, 2011, Social networking sites and our lives. Retrieved July 12, 2011 from.
[20] P. Holme, B.J. Kim, Growing scale-free networks with tunable clustering, Physical Rev. E 65 (2) (2002) 026107.
[21] T. Hossmann, F. Legendre, G. Nomikos, T. Spyropoulos, Stumbl: Using facebook to collect rich datasets for opportunistic networking research, in: IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE, 2011, pp. 1–6.
[22] L. Jian-Guo, D. Yan-Zhong, W. Zhong-Tuo, Multistage random growing small-world networks with power-law degree distribution, Chinese Phys. Lett. 23 (3) (2006) 746.
[23] B. Kantarci, V. Labatut, Classification of complex networks based on topological properties, in: Proceedings of the 3rd International Conference on Social Computing and Its Applications, 2013.
[24] P. Klimek, S. Thurner, Triadic closure dynamics drives scaling laws in social multiplex networks, New J. Phys. 15 (6) (2013) 063008.
[25] J. Leskovec, Stanford large network dataset collection, 2011, URL http://snap.stanford.edu/data/index.html.
[26] J. Leskovec, J.J. Mcauley, Learning to discover social circles in ego networks, Adv. Neural Informat. Process. Syst. (2012) 539–547.
[27] W. Li, J.-Y. Yang, Comparing networks from a data analysis perspective, in: Complex Sciences, Springer, 2009, pp. 1907–1916.
[28] Y. Li, X. Qian, D. Wang, Extended HK evolving network model, in: 24th Chinese Control and Decision Conference (CCDC), 2012, IEEE, 2012, pp. 4095–4097.
[29] I. Lunden, 73 twitter in popularity, facebook stays on top, 2013, techcrunch.com.
[30] A. Masoudi-Nejad, F. Schreiber, Z. Kashani, Building blocks of biological networks: a review on major network motif discovery algorithms, IET Syst. Biol. 6 (5) (2012) 164–174.
[31] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (11) (1995) 39–41.
[32] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of evolved and designed networks, Science 303 (5663) (2004) 1538–1542.
[33] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[34] M.E. Newman, The structure of scientific collaboration networks, in: Proceedings of the National Academy of Sciences, vol. 98, 2001, pp. 404–409.
[35] D. Papo, J.M. Buldú, S. Boccaletti, E.T. Bullmore, Complex network theory and the brain, Philosophical Trans. Royal Soc. B: Biol. Sci. 369 (1653) (2014) 20130520.
[36] P. Parigi, L. Sartori, The political party as a network of cleavages: Disclosing the inner structure of italian political parties in the seventies, Social Netw. (2012).
[37] K. Park, Y. Han, Y.-K. Lee, An efficient method for computing similarity between frequent subgraphs, in: Proceedings of the 3rd International Conference on Social Computing and Its Applications, 2013.
[38] M.Q. Pasta, Z. Jan, A. Sallaberry, F. Zaidi, Tunable and growing network generation model with community structures, in: Third International Conference on Cloud and Green Computing (CGC), 2013. IEEE, 2013, pp. 233–240.
[39] B. Rieder, Studying facebook via data extraction: the netvizz application, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 346–355.
[40] C. Scholz, M. Atzmueller, M. Kibanov, G. Stumme, Predictability of evolving contacts and triadic closure in human face-to-face proximity networks, Soc. Netw. Anal. Mining 4 (1) (2014) 1–17.
[41] C. Song, S. Havlin, H.A. Makse, Self-similarity of complex networks, Nature 433 (7024) (2005) 392–395.
[42] S.H. Strogatz, Exploring complex networks, Nature 410 (6825) (2001) 268–276.
[43] C.-Y. Teng, Y.-R. Lin, L.A. Adamic, Recipe recommendation using ingredient networks, in: Proceedings of the 3rd Annual ACM Web Science Conference, ACM, 2012, pp. 298–307.
[44] A. Topirceanu, M. Udrescu, Measuring realism of social network models using network motifs, in: IEEE 10th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2015. IEEE, 2015.
[45] A. Topirceanu, M. Udrescu, M. Vladutiu, Network fidelity: A metric to quantify the similarity and realism of complex networks, in: Proceedings of the 3rd International Conference on Social Computing and Its Applications, 2013.
[46] A. Topirceanu, M. Udrescu, M. Vladutiu, Genetically optimized realistic social network topology inspired by facebook, in: Online Social Media Analysis and Visualization, Springer, 2014, pp. 163–179.
[47] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabasi, Human mobility, social ties, and link prediction, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 1100–1108.
[48] J. Wang, L. Rong, Evolving small-world networks based on the modified ba model, in: International Conference on Computer Science and Information Technology, 2008. ICCSIT'08. . IEEE, 2008, pp. 143–146.
[49] P. Wang, J. Lü, X. Yu, Identification of important nodes in directed biological networks: A network motif approach, PloS one 9 (8) (2014) e106132.
[50] X.F. Wang, G. Chen, Complex networks: small-world, scale-free and beyond, in: Circuits and Systems Magazine, vol. 3, IEEE, 2003, pp. 6–20.
[51] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (6684) (1998) 440–442.
[52] S. Wernicke, Efficient detection of network motifs, IEEE/ACM Trans. Comput. Biol. Bioinformat. 3 (4) (2006) 347–359.
[53] S. Wernicke, F. Rasche, Fanmod: a tool for fast network motif detection, Bioinformatics 22 (9) (2006) 1152–1153.
[54] S. Wuchty, Z.N. Oltvai, A.-L. Barabási, Evolutionary conservation of motif constituents in the yeast protein interaction network, Nature Genetics 35 (2) (2003) 176–179.
[55] F. Zaidi, Small world networks and clustered small world networks with random connectivity, Soc. Netw. Anal. Mining (2013) 1–13.