To Stay or Not to Stay: Modeling Engagement Dynamics in Social Graphs

Fragkiskos D. Malliaros[1], Michalis Vazirgiannis[1,2]
[1] Computer Science Laboratory, École Polytechnique, France
[2] Department of Informatics, Athens University of Economics and Business, Greece
{fmalliaros, mvazirg}@lix.polytechnique.fr

ABSTRACT

Given a large social graph, how can we model the engagement properties of nodes? Can we quantify engagement both at node level as well as at graph level? Typically, *engagement* refers to the degree that an individual participates (or is encouraged to participate) in a community and is closely related to the important property of nodes' departure dynamics, i.e., the tendency of individuals to leave the community. In this paper, we build upon recent work in the field of game theory, where the behavior of individuals (nodes) is modeled by a technology adoption game. That is, the decision of a node to remain engaged in the graph is affected by the decision of its neighbors, and the "best practice" for each individual is captured by its *core number* – as arises from the k-core decomposition. After modeling and defining the engagement dynamics at node and graph level, we examine whether they depend on structural and topological features of the graph. We perform experiments on a multitude of real graphs, observing interesting connections with other graph characteristics, as well as a clear deviation from the corresponding behavior of random graphs. Furthermore, similar to the well known results about the robustness of real graphs under random and targeted node removals, we discuss the implications of our findings on a special case of robustness – regarding random and targeted node departures based on their engagement level.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data mining*

General Terms

Theory, Experimentation, Measurement

Keywords

Social Engagement; Social Network Analysis; Graph Mining

1. INTRODUCTION

Over the last years, there is a considerable interest on studying the properties and dynamics of social networks, arising from a plethora of online social networking and social media applications, such as FACEBOOK, GOOGLE+ and YOUTUBE. Typically, users become members of an online community for several reasons (e.g., create new friendship relationships, use of applications offered by a platform, etc.) and a lot of research effort has been devoted to understand the dynamics of formation and evolution of those social communities. Characteristic example is the observation that individuals decide to join a community based not only on the number of friends that are already part of the community, but also on the degree of interactions among these friends [5].

However, similar to the decision of becoming member of a community, an individual may also decide to leave the network. Although in many of the popular social networking applications typically users do not explicitly leave the network, the decision of departure can be expressed by inactivity, i.e., the user do not participate in the activities of the community. Can we model and quantify the departure dynamics of individuals in a social graph?

In this paper, we are trying to answer the above question studying the property of *user engagement* in social interaction graphs. Typically, user engagement refers to the extend that an individual is encouraged to participate in the activities of a community[1]. In the areas of sociology and economics, the problem of social engagement examines the engagement of individuals to products, services or ideas. Similarly, in the field of web mining, the property of engagement refers to the quality of the user experience, as expressed by the duration and frequency that a web application is used [6]. In the context of a social graph, the property of engagement captures the *incentive* of a user (node) to remain engaged. In other words, the property of node engagement can be considered as complementary to the one of node departure.

Typically, an individual decides to remain engaged in the community (instead of depart), based on the benefit that is derived by the participation. Intuitively, the benefit of a user is based on its neighborhood, i.e., the number of friends that are also part of the community. Furthermore, as mentioned earlier, the strength of interactions among the friends of a user, is also a crucial factor for being part of the community. Therefore, it becomes clear that the decision of a user to remain engaged is affected by the structure of its neighborhood. Suppose now that a user decides to dropout, due to its low incentive of being part of the community. This decision is possible to affect the engagement level of his neighbors, that potentially can depart as well. This effect can evolve in a *contagion* within the graph, leading to a *cascade* of node departures.

In this paper, we model and study the engagement properties of real-world social graphs. Our approach capitalizes on recent results in the field of game theory, where the engagement property can be considered in a similar manner as a product adoption process

[1] http://en.wikipedia.org/wiki/Social_engagement .

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505561.

[17, 12, 8]. In the case where individuals decide simultaneously whether to remain engaged or depart from the graph, the engagement level of each node can be captured by the properties of the k-core decomposition [22]. Based on this point, we propose measures for characterizing the engagement at both *node level* as well as at *graph level*. We examine in detail the properties of a large number of real graphs, trying to better understand the engagement dynamics.

The main contributions of the paper can be summarized as follows:

- *Problem statement:* We study the property of engagement in social graphs and how it can be used to model the departure dynamics of nodes in the graph.
- *Measures of engagement:* Based on game theoretic models, we propose interesting measures for characterizing the engagement at both node level as well as at graph level.
- *Experiments on real graphs:* We perform a large number of experiments in several real-world graphs, examining the engagement dynamics and observing interesting properties.
- *Implications of our study:* We discuss the implications of our study regarding a new problem of robustness/vulnerability assessment in real graphs, where nodes decide to leave the graph based on their own incentives.

The rest of the paper is organized as follows. Section 2 gives the related work and Section 3 provides the necessary background. Then, in Section 4 we describe the model and the proposed engagement measures. Section 5 presents the experimental evaluation of our method, while in Section 6 interesting implications of our study are discussed. Finally, we conclude in Section 7.

2. RELATED WORK

In this section we review the related work, regarding the engagement dynamics in social graphs, as well as other applications of the k-core decomposition. In the very recent literature, there has been presented some game-theoretic models for the problem of product adoption in networked environments [17, 12, 8]. These models form the basis of our approach and are described in detail in Section 4. To the best of our knowledge, the only related work that provides experimental study for the problem of node departures in social networks, is the work presented in [25]. There, the authors study two real social networks and examine whether the departure dynamics show similar behavior with the arrival dynamics. In the case of node departures, the authors of [25] observed that the active users typically belong to a dense core of the graph, while inactive users are placed on the sparsely connected periphery of the graph. As we will present later, the property of node engagement can be considered complementary to the one of node departure, and our approach provides a more refined modeling of the observations made in [25]. Moreover, related to our work can be considered recent studies about the formation dynamics of communities [5, 14], as well as studies about diffusion and contagion in social graphs [23, 4, 21, 10].

As we will present in Section 4, our method builds upon the properties of the k-core decomposition in a graph (see Section 3 for more details). Broadly speaking, the k-core decomposition has been applied in the past for extracting the most coherent subgraphs [22], graph visualization [26], identification of influential spreaders [13], and for studying [3] and modeling [9] the Internet topology. In this work, we examine one more application domain of the k-core decomposition in the problem of node engagement in social graphs; in contrast to the previous studies that mostly focus on the nodes of the best k-core subgraph, in this paper we are interested in the hierarchy produced by the decomposition, since it can provide meaningful insights about the engagement dynamics of the graph.

3. PRELIMINARIES AND BACKGROUND

In this section we briefly discuss the properties of the k-core decomposition [22], which is utilized by our method. Let $G = (V, E)$ be an undirected graph, where $|V| = n$ and $|E| = m$. A graph H is a subgraph of G, denoted by $H \subseteq G$, if H can be obtained from G after removing edges or vertices.

DEFINITION 1 (k-CORE SUBGRAPH). *Let H be a subgraph of G, i.e., $H \subseteq G$. Subgraph H is defined to be a k-core of G if it is a maximal connected subgraph of G, in which all nodes have degree at least k.*

DEFINITION 2 (GRAPH DEGENERACY $\delta^*(G)$). *The degeneracy $\delta^*(G)$ of a graph G is defined as the maximum k for which graph G contains a non-empty k-core subgraph.*

DEFINITION 3 (NODE'S CORE NUMBER). *A node i has core number $c_i = k$, if it belongs to a k-core but not to any $(k+1)$-core.*

A k-core of a graph G can be obtained by repeatedly deleting all vertices of degree less than k. Furthermore, the k-core decomposition – which assigns a core number c_i to each node $i \in V$ – can be computed efficiently, with complexity $\mathcal{O}(m + n)$ proportional to the size of the graph [7]. The most important point is that the k-core decomposition creates an hierarchy of the graph, where "better" k-core subgraphs (i.e., higher values of k) correspond to more cohesive parts of the graph.

4. PROBLEM FORMULATION AND PROPOSED METHOD

In this section, we formulate the problem of modeling and quantifying the engagement dynamics in a social interaction graph. We begin by discussing the main factors that intuitively affect the decision of nodes to remain engaged or leave the graph. Then, we present the theoretical model used to approximate and capture the engagement behavior of nodes, as well as the proposed engagement measures at both node and graph level.

4.1 Problem Statement and Model Description

Our goal is to model and study the problem of node engagement in social graphs. Informally, the property of engagement captures the incentive of individuals to remain engaged in the graph, as opposed to their decision of departure. In the context of this paper, we are interested in the engagement dynamics of individuals as well as of the whole system, from a *network-wise* point of view. In other words, we consider only the underlying graph structure of a social system, and based on its properties we derive measures that characterize the behavior in terms of engagement.

Typically, each individual that participates in a social activity – as expressed by his/her participation in a social graph – derive a benefit. In most of the cases, this benefit emanates from his/her neighborhood, as captured by the node degree in the social graph. Furthermore, one additional factor that affects the benefit of each individual is the degree of interaction among its neighbors [23], in the sense that if one's friends tend to highly interact among each other, the benefit of remaining engaged in the graph could potentially be increased.

Let us now suppose that a user decides to drop out from his community due to the fact that the incentive of staying has been reduced. This decision will cause direct effects in his neighborhood, in the sense that some of his friends may also decide to depart. More precisely, a departure can become an *epidemic* (or contagion), forming a *cascade* of individual departures; nodes will decide to leave and this will also affect not only their neighbors but also the whole community. Therefore, according to the notion of *direct-benefit effects*, individuals who want to incur an explicit benefit by remaining engaged, they should align their decision with the one of their neighbors [10].

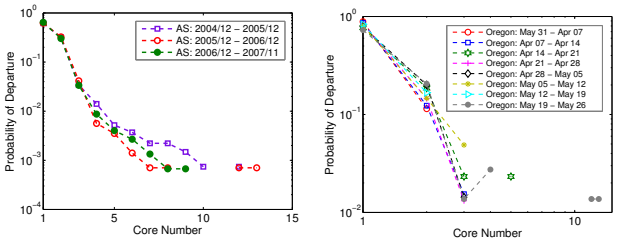
Next, we present our model and the proposed measures for engagement in social graphs. Each node $v \in V$ – that corresponds to an individual – can either remain engaged in the network or can decide to depart. As we mentioned earlier, it is natural that the decision of each node should be based on the decisions of its neighbors. The behavior of nodes as a system can be expressed using game-theoretic concepts, and more precisely it can be captured by the notion of *networked coordination games* [10]. That is, the property of engagement can be viewed as a network model based on *direct-benefit* effects: the node's benefit of remaining engaged in the graph increases as more neighbors decide respectively to stay in the graph. This formulation has been extensively studied in the areas of game theory and economics. It is applied in situations where the nodes have to choose between two possible alternatives and the structure of the underlying social network affects the decision: for two neighborhood nodes u and v, there is an incentive to be aligned with the same decision, since that way they will both increase their benefits produced by the underlying interactions.

In a similar way, since the benefit of each node for staying in the network emanates from its neighbors, the problem can be modeled using similar concepts with the ones of coordination games. We consider that the nodes of the graph – which correspond to rational individuals/players – decide *simultaneously* whether to stay or leave. Each node $i \in V$ has the same set of possible strategies $\mathcal{X} = \{0, 1\}$, i.e., *leave* or *stay* in our case. Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ be the vector that denotes the decision of each node. The *payoff* (or utility) of a node i given the behavior of the rest nodes (as captured by vector $\mathbf{x}$), can be expressed as:

$$\Pi_i(\mathbf{x}) = \textbf{benefit}\Big(x_i, \sum_{j \in \mathcal{N}_i} x_j\Big) - \textbf{cost}(x_i), \qquad (1)$$

where **benefit**$(\cdot)$ and **cost**$(\cdot)$ are the node's benefit and cost functions respectively and $\mathcal{N}_i = \{j \in V : (i, j) \in E\}$ is the neighborhood set of node i. In other words, the benefit of each node depends on its own decision x_i and the aggregate decisions of its neighbors; this captures at a large extend the problem of engagement estimation, since in many cases a user remains engaged according to the degree of interactions with its friends in the community. Furthermore, every node i incurs a cost for remaining engaged in the graph, which depends only on its own action. While the actual form of the cost function does not need to be apriori known in the model, it is clear that a node will decide to stay engaged if its cost is not higher than its benefit. Let **cost**$(x_i) = k$ be the cost value of each node $i \in V$. Then, according to Eq. (1), every node that will remain engaged should have non negative payoff, and therefore $\Pi_i(\mathbf{x}) = |\mathcal{N}_i^{x=1}| - k$, where $|\mathcal{N}_i^{x=1}|$ is the number of i's neighbors that finally remain engaged, i.e., the degree of i in the graph induced by nodes with decision $x = 1$.

Thanks to some very recent results in the area of game theory, the equilibrium of this game corresponds to the core number c_i of each node, as produced by the k-core decomposition [17, 12, 8].



(a) CAIDA (b) OREGON

Figure 1: Probability of departure vs. core number. The discontinuities in the plot correspond to zero (we plot only the cores from which nodes depart).

PROPOSITION 1 (EQUILIBRIUM PROPERTY, [17, 12]). *The best response (Nash equilibrium) of each node $i \in V$ in the model presented above corresponds to the core number c_i.*

In other words, in the case of equilibrium, every node in the induced subgraph S formed by nodes with $x_i = 1$ should have minimum degree k, satisfying the property of $c_i \geq k$. That way, no engaged node will have incentive to depart from S and no node outside S, i.e., in $V - S$, will have at least k neighbors in S in order to remain engaged after his departure. As noted by the authors in [17, 12, 8], the game has multiple equilibria, but the maximum one corresponds to the "best" k-core structure of the graph (i.e., $k = \delta^*(G)$). In our case, we are interested in all nodes in the graph and not only on those that form the best equilibrium; as we will present shortly these nodes show interesting properties.

4.2 Engagement Measures

Having presented the basic theoretical model, we now proceed with the proposed measures for characterizing the engagement dynamics on graphs. We are interested in studying the engagement properties at both *node* (local) and *graph* (global) level; furthermore, we are interested in examining the behavior of specific subgraphs – as produced by the k-core decomposition – which include nodes with specific engagement level.

Capitalizing on Proposition 1, we quantify the property of node engagement using the k-core decomposition, and more specifically the core number c_i of each node $i \in V$.

PROPOSITION 2 (NODE ENGAGEMENT). *The engagement level e_i of each node $i \in V$ is defined as its core number c_i.*

Typically, nodes that belong to higher cores of the graph (higher core number), show better engagement and therefore it is less probable to depart (or, at least, the incentive to depart is lower). As we discussed in Section 4.1, the core number of each node is a reasonable metric (or estimator) to capture and model the engagement dynamics: nodes remain engaged if their neighbors (and the neighbors of their neighbors, etc.) also remain engaged. On the other hand, if a node decides to depart, this may affect the engagement level of its neighbors – which may decide to leave as well – forming a cascade of potential departures in the graph. This dynamic effect of cascades is naturally captured by the k-core decomposition and the core number of nodes.

Figure 1 provides some empirical observations regarding the departure of nodes – on data with available relevant information – that can be used to support our modeling approach. As it is difficult to access social graph data where nodes (users) explicitly define their departure time, we have examined two snapshots of the Internet

topology (CAIDA and OREGON Autonomous Systems) with available dropout information. Figure 1 depicts the probability of departure vs. the core number c_i of a node. As it can be observed, nodes that belong to smaller cores (close to the first core) are more probable to leave the graph, thus supporting our modeling approach. Moreover, this property is persistent for several time snapshots of the graphs.

Additionally, the proposed engagement metric can be considered as a more refined modeling explanation of the departure dynamics in social graphs, as very recently observed in [25]. The authors of Ref. [25] studied the behavior of user departures in social networks and based on some inactivity criteria (e.g., in a co-authorship network, a user is considered inactive if he/she has not published a paper in a time period of more than five years), they observed that nodes which belong to the densely connected core of the graph mainly correspond to active users. On the other hand, inactive users (i.e., users that potentially have left the graph) belong to the periphery of the graph. In other words, the departure of nodes is proportional to their position in the graph, with nodes in the fringe of the graph presenting higher probability to dropout. Our modeling approach and the node engagement metric e – based on the properties of the k-core decomposition – quantify in a precise manner the above structural observations.

We should note that, here we examine the dynamics of engagement (and thus of departure) by a simple model and metric, that approximates real settings and observations in a concise manner. However, we do not argue that the engagement of a user is solely proportional to his core number; other external factors may affect its behavior as well. Nevertheless, in the rather realistic case where each node decides to remain engaged for maximizing its revenue by the participation in the community – thus considering the decision of its neighbors – the behavior can be modeled by the proposed metric.

Furthermore, as we will see in the experimental results, the degree of a node is not an accurate estimator of the departure dynamics: while high degree is necessary for achieving higher engagement and higher core number, the opposite is not always true. In many cases, high degree nodes have low core number because of the fact that their neighbors are not well connected among each other[2]. Therefore, the engagement should be described by a metric able to capture both the size of node's neighborhood as well as its connectivity. In Section 5 where the experimental results are presented, we also examine how other well-known structural characteristics of the graph (e.g., degree, triangle participation ratio, clustering coefficient) affect the engagement behavior.

Based on this model of user behavior, we also propose to study the characteristics of the subgraphs produced in the case of simple scenarios, where nodes with certain engagement index k (for various values of k) decide simultaneously to drop out. The subgraph that remains after such types of departures is defined as the k-engagement subgraph $\mathcal{G}_k$.

DEFINITION 4 (k-ENGAGEMENT SUBGRAPH $\mathcal{G}_k$). *Let k be a integer parameter such that a node remains engaged in G if at least k neighbors are engaged. The graph $\mathcal{G}_k$ which is induced by nodes $i \in V$ with engagement level $e_i \geq k$ is defined as the k-engagement subgraph.*

The k-engagement subgraphs correspond to interesting structures of the graph. Actually, for a specific value of k, subgraph $\mathcal{G}_k$ represents the remaining graph, after the cascading effect where nodes with engagement lower than k have left the graph. The properties of the remaining subgraph – as captured by $\mathcal{G}_k$ – are crucial towards a better understanding of the engagement characteristics, as well as for examining the functional operation of the remaining graph. As we will present shortly, the size distribution of $\mathcal{G}_k$ for various values of k can inform us about the overall engagement level of the graph. Furthermore, it is interested to study whether well-known structural properties – such as the degree distribution of the graph – are retained after such types of nodes' departures. We also note that, following the properties of k-core decomposition, subgraphs $\mathcal{G}_k$ form a nested hierarchy $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \mathcal{G}_2 \supseteq \ldots \supseteq \mathcal{G}_k$ for the possible values of k, in the sense that subgraphs of higher k also belong to $\mathcal{G}_k$'s of lower k.

Of particular interest is the subgraph $\mathcal{G}_k$ that corresponds to the maximum value of engagement e. In terms of k-core decomposition, the nodes with the highest engagement level $e_{\max}$ are those who belong to the best k-core of the graph, i.e., $k = \delta^*(G)$, where $\delta^*(G)$ is the degeneracy of the graph [22].

PROPOSITION 3 (MAX-ENGAGEMENT SUBGRAPH). *Let $k = \delta^*(G)$ be the degeneracy of the graph, i.e., the maximum k such that there exists a k-engagement subgraph. In our context, we consider this value as the maximum engagement level of the graph, i.e., $e_{\max} = \delta^*(G)$ and we denote the Max-Engagement Subgraph as $\mathcal{G}_{e_{\max}}$.*

The Max-Engagement subgraph is composed by the nodes of the graph that show the highest engagement level $e = e_{\max}$. More precisely, each node $i \in \mathcal{G}_{e_{\max}}$ has degree $d_i \geq e_{\max}$ within $\mathcal{G}_{e_{\max}}$, implying that this set of nodes has potentially the lowest incentive to depart from the graph and thus it corresponds to the best engaged nodes. As we will discuss in Section 6, this subgraph also contains the most influential nodes of the graph, in terms of departure dynamics.

Having defined the engagement index of each node in the graph as well as the notion of the k-engagement subgraphs, it would be interesting to summarize this information into one value capable to describe the engagement level of the whole graph. That is, each individual node contributes to the engagement of the graph – according to its engagement index $e_i, \forall i \in V$ – based on the best core c_i that the node belongs to. Ideally, in terms of engagement, it would be better to have a large fraction of the nodes of the graph belonging to largest cores, thus showing higher engagement. In the extreme case of the graph with the *best engagement properties* – which corresponds to the complete graph $K_n = (V_{K_n}, E_{K_n})$ – all nodes belong to the Max-Engagement subgraph and their engagement index is equal to $e_i = |V_{K_n}| - 1, \forall i \in V_{K_n}$. In order to capture this behavior, we consider the *area under the curve* of the cumulative distribution of the k-engagement subgraphs' sizes. However, since the graphs do not have the same maximum engagement level $e_{\max}$, we normalize this value for each graph into the interval $[0, 1]$, based on a simple normalization: Normalized $e_j = \dfrac{e_j - \min(e)}{\max(e) - \min(e)}$, where $j = 1, \ldots, e_{\max}$, $\min(e) = 1$ and $\max(e) = e_{\max}$ (for simplicity, we consider that all nodes in the graph have degree at least one, therefore the minimum engagement is 1).

PROPOSITION 4 (GRAPH ENGAGEMENT). *Let $\mathcal{F}(e) = \Pr(X \geq e)$ be the cumulative distribution function of the sizes of the k-engagement subgraphs. Then, the total engagement level of a graph G, denoted as $\mathcal{E}_G$, is defined as the area under the curve of $\mathcal{F}(e)$, $e = [0, 1]$, i.e.,*

$$\mathcal{E}_G = \int_0^1 \mathcal{F}(e)\, de. \tag{2}$$

[2]A similar behavior has been reported in [3] in the context of Internet graph analysis.

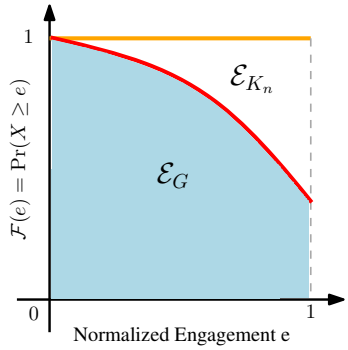


Figure 2: Schematic representation of the engagement index $\mathcal{E}_G$. The red curve shows an example of the cumulative distribution of the engagement level of the k-engagement subgraphs. The area under the curve (light blue region) captures the engagement properties of the graph. The orange colored curve shows the engagement level of the complete graph K_n.

Figure 2 depicts a schematic representation of the engagement index $\mathcal{E}_G$. The horizontal axis corresponds to the normalized engagement value e, while the vertical axis represents the probability $\Pr(X \geq e)$ that a node has (normalized) engagement level at least e (as produced by the sizes the k-engagement subgraphs). The values of $\mathcal{E}_G$ are in the range of $[0, 1]$, with higher values indicating graphs with higher total engagement level (larger area under red curve). In the case of the complete graph (orange colored curve), the probability that a node has normalized engagement at least $e, \forall e \in [0, 1]$, is $\Pr(X \geq e) = 1$. In other words, every node in the graph has engagement $e_i = |V_{K_n}| - 1$, and therefore the size of the k-engagement subgraphs, for $k = 1, \ldots, e_{\max}$, is equal to the size of the whole graph, i.e., $|V_{K_n}|$.

4.3 Discussion

Having presented the proposed engagement measures, we briefly discuss on an important point in the modeling approach followed by our method. Our approach and the proposed engagement measures are build upon the game presented in Section 4.1, which considers that nodes have *complete information* about the structure of the graph [10, 17]. Although this assumption may not be very accurate in many settings where individuals should take a decision (to remain engaged or to depart in our problem), we consider that in this case is valid since our goal is to model and to provide a high level study of the behavior of individual nodes and of the graph as a whole, regarding their engagement properties. Thus, our study builds upon the fact that we have knowledge of the structure of the graph. In a typical application scenario of our approach, the administrator of a social graph (e.g., FACEBOOK) – who has global knowledge of the structure of the graph – can use the proposed measures to examine the engagement dynamics of the graph and to potentially detect nodes that tend to leave, due to their low engagement.

5. ENGAGEMENT OF REAL GRAPHS

In this section we present detailed experimental results of the proposed engagement measures, at both local (node) and global (graph) level. The experiments were designed to address the following points:

P1: Study the characteristics of the engagement dynamics in real graphs.

P2: Examine how other graph features affect the engagement of the graph.

As we have already mentioned, we consider that the feasibility and applicability of our approach is supported by the results depicted in Fig. 1 and by very recent observations about the departure dynamics in social graphs [25]. Actually, here we present a more refined explanation of the departure dynamics, studying the complementary property of engagement. Furthermore, the time complexity of our approach is linear with respect to the size of the graph, as it relies on the computation of the k-core decomposition (see Section 3).

Table 1 presents the datasets used in our study. All of them are publicly available and correspond to well-known social and collaboration networks (except from the last datasets used to support our modeling approach).

5.1 High Level Properties of k-Engagement Subgraphs

As we described in Section 4, a reasonable estimator for the engagement properties of a node is its core number, i.e., $e_i = c_i, \forall i \in V$. One important aspect here is to examine the size distribution of the k-engagement subgraphs, i.e., the size of the subgraphs that contain nodes with engagement e at least k. That is, for the various possible values of parameter k (that depend on the graph), we study the properties of the k-engagement subgraphs. These characteristics can help us to further understand the engagement dynamics of real graphs, both at node and at graph level. Figure 3 (red curve) depicts the results for the real graphs presented in Table 1.

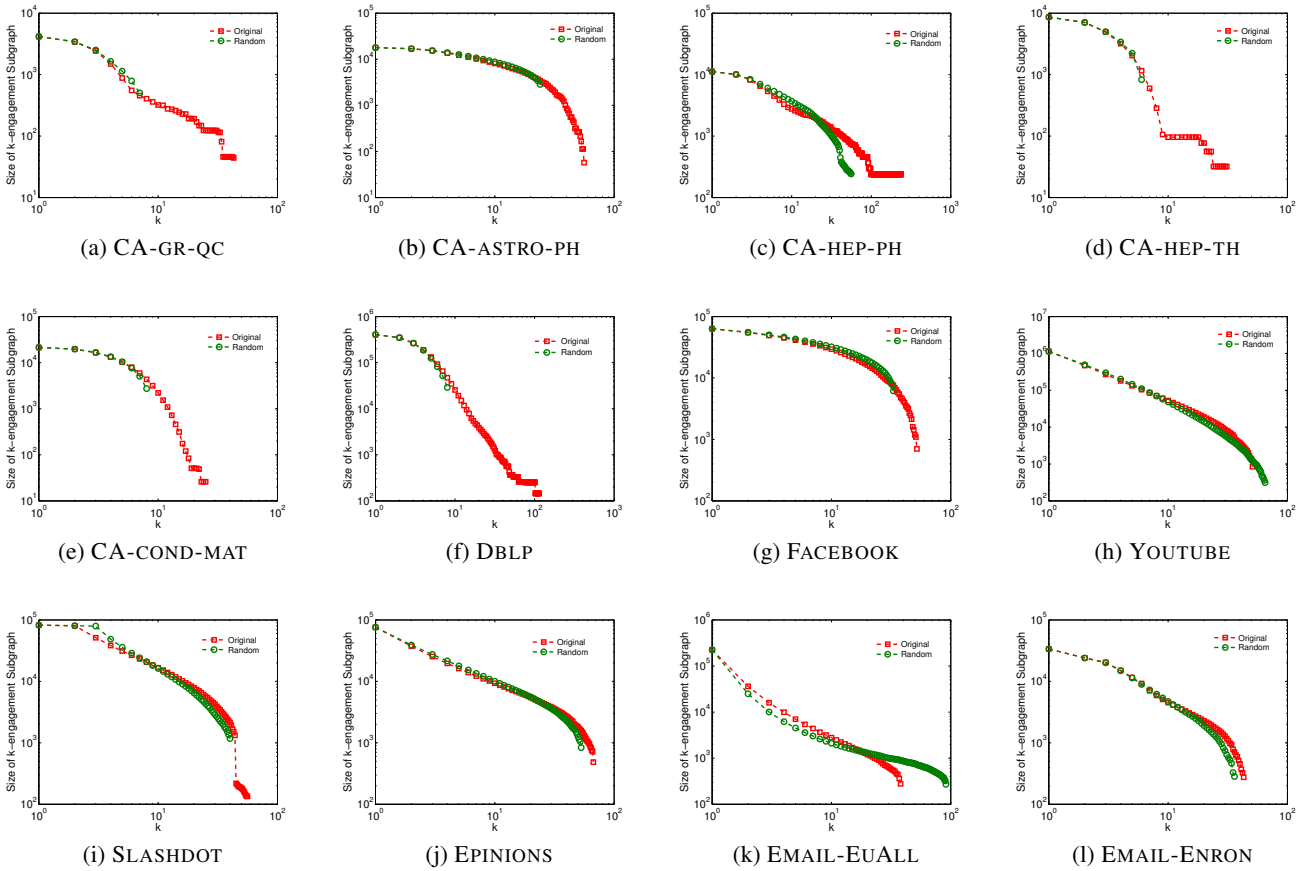
As we can observe, for most of the datasets, the distribution of the sizes of the k-engagement subgraphs is almost skewed, meaning that the highly engagement subgraphs (for larger values of k) are relatively small in size. In other words, most of the nodes in the graph have small engagement index e, while a few nodes are highly engaged. Of course, we should note that the size distributions are not identical for all the graphs we have examined. Furthermore, the maximum engagement level $e_{\max}$ as well as the size of the Max-Engagement subgraph $\mathcal{G}_{\max}$ – that corresponds to the tail of the distribution – present different behavior for some of the examined datasets. We will discuss these points next in the paper.

One important question here is if these observations regarding the engagement properties of graphs, capture the behavior of a real system – and thus can be characterized as patterns of real graphs. In other words, is there any difference between the $\mathcal{G}_k$'s size distribution of real graphs and random ones? To answer this question, we have examined the engagement properties of a *configuration model*, i.e., a random graph model with the same degree distribution as the original one. As we can observe from Fig. 3 (green curve), a random rewiring of the original graph causes a different size distribution for the k-engagement subgraphs. More precisely, for most of the examined datasets, the random equivalent graph shows a much lower number of engagement levels, but the size of the Max-Engagement subgraph is much larger compared to the original one – indicating different behavior in terms of engagement. This observation is somewhat expected; random graphs are known to have a large core and thus Max-Engagement subgraphs of relatively large size [19].

However, for a few datasets we have observed an unexpected but rather interesting behavior. For example, YOUTUBE and EMAIL-EUALL social graphs show an almost similar size distribution between the original graph and the random equivalent one. Additionally, EMAIL-EUALL has a much smaller maximum engagement index $e_{\max}$ compared to the random rewired graph. To better examine this deviation as well as for having a more refined explanation

Table 1: Summary of real-world networks used in this study.

Network Name	Nodes	Edges	Description
FACEBOOK [24]	63,392	816,886	Facebook New Orleans social network
YOUTUBE [18]	1,134,890	2,987,624	Social network from Youtube
SLASHDOT [16]	82,168	582,533	Slashdot social network (Feb. '09)
EPINIONS [20]	75,877	405,739	Who trusts whom network
EMAIL-EUALL [15]	224,832	340,795	E-mail network
EMAIL-ENRON [16]	33,696	180,811	E-mail network
CA-GR-QC [15]	4,158	13,428	Co-authorship network in General Relativity
CA-ASTRO-PH [15]	17,903	197,031	Co-authorship network in Astro Phys.
CA-HEP-PH [15]	11,204	117,649	Co-authorship network in High Energy Phys.
CA-HEP-TH [15]	8,638	24,827	Co-authorship network in High Energy Phys. Th.
CA-COND-MAT [15]	21,363	91,342	Co-authorship network in Condensed Matter
DBLP [1]	404,892	1,422,263	Co-authorship network from DBLP
CAIDA/OREGON [15]	26,475/11,461	106,762/32,730	Autonomous systems graphs



(a) CA-GR-QC (b) CA-ASTRO-PH (c) CA-HEP-PH (d) CA-HEP-TH

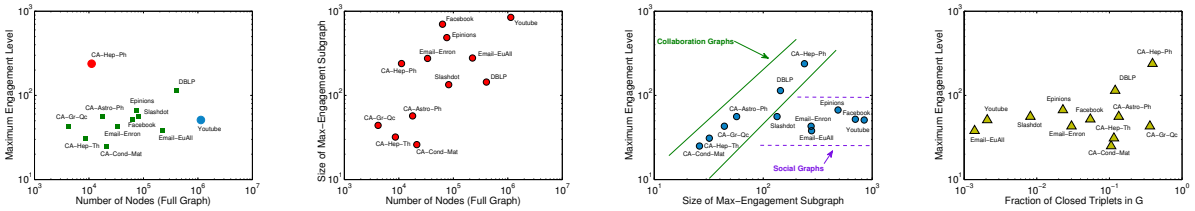
(e) CA-COND-MAT (f) DBLP (g) FACEBOOK (h) YOUTUBE

(i) SLASHDOT (j) EPINIONS (k) EMAIL-EUALL (l) EMAIL-ENRON

Figure 3: Size distribution of the k-engagement subgraphs. Each plot depicts the size distribution of the k-engagement subgraphs vs. k, where $k = 1, \ldots, e_{\max}$. The red line corresponds to the distribution of the original graph, while the green one to the random graph with the same degree sequence as the original one.

of the observed engagement properties, we have computed a set of high level characteristics of the k-engagement graphs. More specifically, we focus on the properties of the Max-Engagement subgraph $\mathcal{G}_{e_{\max}}$, trying to understand and capture which factors potentially affect the engagement properties of the graph.

Figure 4 depicts the relationship between some high level characteristics of the Max-Engagement subgraphs $\mathcal{G}_{e_{\max}}$ with important global features of the graph (for the datasets of Table 1). We argue that examining interesting correlations between graph features and the observed engagement characteristics, we can draw meaningful conclusions about the engagement dynamics of real graphs.



(a) $|V|$ vs. $e_{\max}$ (b) $|V|$ vs. # of nodes in $\mathcal{G}_{e_{\max}}$ (c) # of nodes in $\mathcal{G}_{e_{\max}}$ vs. $e_{\max}$ (d) Closed triplets in G vs. $e_{\max}$

Figure 4: Characteristics of the $\mathcal{G}_{e_{\max}}$ subgraphs of the studied graphs (Table 1). (a) Maximum engagement level $e_{\max}$ vs. the size of the whole graph. (b) Number of nodes in the Max-Engagement subgraph vs. the size of the whole graph (Pearson correlation coefficient $\rho = 0.6394$). (c) Number of nodes in the Max-Engagement subgraph vs. maximum engagement level $e_{\max}$. Observe the different behavior between the collaboration (co-authorship) graphs and the social networks from social media applications. (d) Maximum engagement level $e_{\max}$ vs. fraction of closed triplets in the whole graph G.

Initially, we consider the relationship between the size of the full graph, i.e., $|V|$ and the characteristics of the Max-Engagement subgraphs, namely the maximum engagement level $e_{\max}$ and the number of nodes in $\mathcal{G}_{e_{\max}}$. As we can see from Fig. 4 (a), for the majority of the examined datasets (green colored squares), the size $|V|$ of the graph shows an almost linear correlation (in log-log axis) with the maximum engagement level $e_{\max}$. However, YOUTUBE (blue colored circle) and CA-HEP-PH (red colored circle) clearly deviate from this relationship (if we ignore these two graphs, Pearson's correlation coefficient is $\rho = 0.75$). While YOUTUBE corresponds to the largest graph of our collection, its $e_{\max}$ value is relatively small. On the other hand, CA-HEP-PH has a relatively small size, while its maximum engagement level is extremely high. Thus, it seems that to achieve a higher $e_{\max}$ value, the size of the graph is not the only responsible factor. That is, as we have already mentioned, the existence of clustering structures in the graph plays a crucial role for the engagement properties.

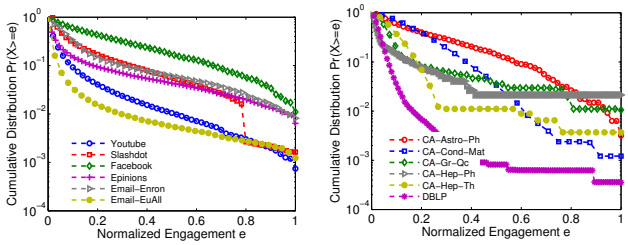
Figure 4 (d) depicts the relationship between the fraction of closed triplets in the graph (triplets of nodes that form triangles) with the maximum engagement level. As we can observe, CA-HEP-PH has the largest fraction of closed triplets in our collection as well as the highest $e_{\max}$ value, although its size is relatively small. On the other hand, YOUTUBE shows an almost opposite behavior. Despite its large size, the fraction of closed triplets and the maximum engagement level are relatively small (in Section 5.4, we present a more detailed examination about the relationship of the engagement and the existence of clustering structures in the graph).

Figure 4 (b) depicts the number of nodes in the graph vs. the number of nodes in the Max-Engagement subgraph $\mathcal{G}_{e_{\max}}$. Here, we can observe a more clear correlation between $|V|$ and the size of $\mathcal{G}_{e_{\max}}$ (Pearson's correlation coefficient $\rho = 0.6394$).

Lastly, in Fig. 4 (c), we study the size of $\mathcal{G}_{e_{\max}}$ vs. the $e_{\max}$ values of the graphs. We can observe two different behaviors in the studied datasets. On the one hand, we have the collaboration graphs – formed by co-authorship relationships. Although they capture different scientific disciplines, we can observe an almost linear correlation in log-log scale. Furthermore, in many cases, the size of $\mathcal{G}_{e_{\max}}$ is almost equal to the maximum engagement level $e_{\max}$, indicating tightly knit communities at this portion of the graph. For example, in the case of the DBLP co-authorship graph, the Max-Engagement subgraph corresponds to a set of around 115 author that have co-authored the same paper. On the other hand, the graphs from online social networking and social media applications, follow a different behavior: the maximum engagement level is kept below a threshold of about 100 and the values are close to each other – almost constant – although the datasets are of different size, while the number of nodes in $\mathcal{G}_{e_{\max}}$ increases. This can be possibly explained by the nature of interactions in online social networking applications; although an individual can achieve a high number of friendship connections, the degree of collaboration – and similarly of engagement – among them is constrained to the threshold of around 100 nodes. We also note that, the value of $e_{\max}$ almost matches the size of the best communities (around 100 nodes) observed by Leskovec et al. [16].

5.2 Graph's Engagement Properties

Having examined the properties of the k-engagement subgraphs, we proceed to the computation of the total graph engagement index $\mathcal{E}_G$. As we described in Section 4, the global engagement properties of the graph can be captured by the area under the curve of the normalized cumulative size distribution of the k-engagement subgraphs. That is, for each graph, we normalize the engagement level e in the interval $[0, 1]$ and we plot the cumulative distribution $\Pr(X \geq e)$, i.e., the fraction of nodes with normalized engagement at least e. Since we are not only interested in the maximum engagement level of each graph but on how individual nodes are distributed within the different levels (as expressed by the k-engagement subgraphs), we are able to compare the engagement properties of different graphs. Figure 5 depicts the results for the collection of graphs of Table 1. Furthermore, Table 2 shows the $\mathcal{E}_G$ values that correspond to the area under curve.



(a) Social graphs (b) Collaboration graphs

Figure 5: Normalized cumulative distribution function of k-engagement subgraphs. Each curve corresponds to the probability $\Pr(X \geq e)$, i.e., the fraction of nodes with normalized engagement at least e. The area under curve captures the engagement index $\mathcal{E}_G$ for the whole graph.

In the case of social graphs (Fig. 5 (a)), we can observe that the graph with the maximum engagement index $\mathcal{E}_G$ is FACEBOOK. Although FACEBOOK does not have a large maximum engagement level $e_{\max}$, nodes are well distributed within levels, with a "good" fraction of nodes having high (normalized) engagement e. Looking

Social Graph	$\mathcal{E}_G$
FACEBOOK	0.2514
YOUTUBE	0.0441
SLASHDOT	0.1221
EPINIONS	0.0755
EMAIL-EUALL	0.0277
EMAIL-ENRON	0.1245

Collab. Graph	$\mathcal{E}_G$
CA-GR-QC	0.0971
CA-ASTRO-PH	0.2293
CA-HEP-PH	0.0651
CA-HEP-TH	0.0969
CA-COND-MAT	0.1924
DBLP	0.0338

Table 2: Graph engagement values $\mathcal{E}_G$ for social (left table) and collaboration (right table) graphs.

now at the collaboration graphs (Fig. 5 (b)), a first observation is that the DBLP graph shows the lower engagement index $\mathcal{E}_G$, compared to the rest co-authorship graphs. One possible explanation is that DBLP covers several areas of computer science, with a significant number of relatively "new" authors. These authors, typically belong to lower cores of the graph, and thus their engagement is relatively low. On the other hand, the rest of the co-authorship networks correspond to more robust communities, where a larger fraction of authors (nodes) has higher engagement level.

5.3 Near Self Similar k-Engagement Subgraphs

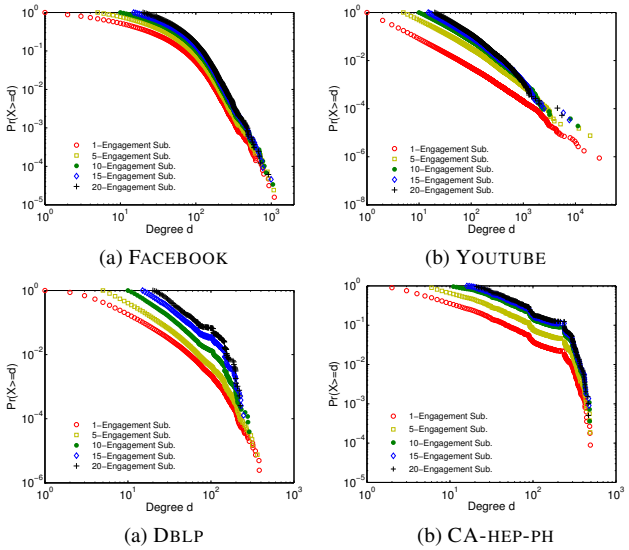
In this section, we are interested to study the properties of the k-engagement subgraphs, under a simple scenario where nodes with low engagement e decide to depart. More specifically, we study the existence of self-similar properties in the k-engagement subgraphs and we focus on the simplest such property which is the existence of a skewed degree distribution. This property is crucial for the k-engagement subgraphs from several viewpoints. First of all, the degree of each node corresponds to an important structural characteristic and therefore it is interesting to examine to what extend it is preserved by the cascade of node departures. Furthermore, the existence of hubs in the k-engagement subgraphs is another crucial point, since – among other things – it is related to how fast information is disseminated in the graph and to the well known type of robustness under targeted/random node attacks [2]. Therefore, we are still interested to examine the major characteristics and functionalities of the graph after a cascade of dropouts.

Figure 6 depicts the cumulative degree distribution of the k-engagement subgraphs, under different values of k (note that, $k = 1$ corresponds to the whole graph). A first observation is that the shape of the distribution is retained for the examined values of k. In other words, for the very first levels of engagement, an almost scale invariance is presented, with respect to the scenario of node departures. However, we do not argue that this property is retained for all the engagement levels, i.e., $e = 1, \ldots, e_{\max}$[3]. Similar properties hold for the rest datasets.

An interesting point here is to examine the diversity of nodes – in terms of degree – that finally depart. In our scenario, nodes with low engagement decide to drop out. How is this mapped to the degree distribution? Typically, we expect that the nodes which depart, correspond to low degree ones. The crucial point here is that the produced cascades can cause the departure of high degree nodes as well, since their engagement level may be reduced. This is actually the major difference between the problem of node departures – based on the engagement level – studied by our paper, compared to removals of nodes based on possible failures [2].

As we can observe from Fig. 6, for different values of k, the tail of the distributions – that captures high degree nodes – is also affected by the cascade of low-engaged node departures. To make

[3]A similar behavior has been also noted for the Internet graph [3].



(a) FACEBOOK (b) YOUTUBE

(a) DBLP (b) CA-HEP-PH

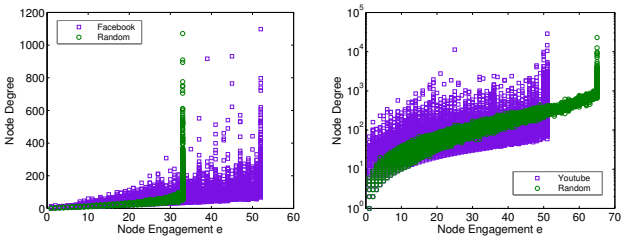
Figure 6: Cumulative degree distribution of k-engagement subgraphs $\mathcal{G}_k$, for various values of k. Note that, the tail of the distribution also changes for different values of k.

this observation more precise, we compute the correlation between nodes' degree and their engagement index e. Clearly, high degree is required to achieve high engagement; however, the degree alone is not an indicator of high engagement. Figure 7 depicts the node engagement index e vs. the degree of each node (we focus only on four datasets of our collection). Clearly, a large fraction of high degree nodes show low engagement level, and thus it is more probable to depart. This is also an indication that the importance of some hub nodes in the graph diminishes, in terms of engagement dynamics. Lastly, we have examined this aspect in the case of random graphs with the same degree distribution as the original ones. As it can be shown from Fig. 7, these graphs show more smooth behavior compared to real ones; this is one more evidence about the differences in terms of engagement between real and random rewired graphs.

5.4 Engagement and Clustering Structures

As we have already discussed, it is expected that the engagement level of a node should be closely related to local clustering structures of the graph, indicating increased level of collaboration among nodes of the same neighborhood. Actually, the authors of [25] report that the probability of departure is related to the overall neighborhood activity of a node. In the more general problem of influence and product/behavior/idea adoption in a social system, the probability that a user will finally proceed with the adoption, is proportional to the size, as well as to the connectivity of the neighborhood [21, 5]. Furthermore, the high level characteristics presented in Fig. 4 (d), indicate a relationship between the fraction of closed triplets in the graph and the maximum engagement level. This is an interesting evidence, in the sense that in higher order k-engagement subgraphs (i.e., higher values of k), a higher degree of cohesiveness exists.

In fact, the number of triangles that each node participates to, seems to be vital for its core number and therefore for its engagement index. Additionally, the fraction of closed triplets in a graph is closely related to the *clustering coefficient* – a measure that inform us about the tendency of nodes to cluster together, forming



(a) FACEBOOK (b) YOUTUBE

(a) DBLP (b) CA-HEP-PH

Figure 7: Node engagement vs. node degree. Correlation between engagement and degree for each node in the graph. Purple squares correspond to real graphs and green circles to the random ones. For the real graphs, observe that high degree nodes can also have relatively small engagement.

tightly knit groups[4]. Recently, Gleich and Seshadhri [11] showed that the number of cores in real-world graphs depends on the global clustering coefficient, where graphs with higher global clustering coefficient tend to have larger number of cores.

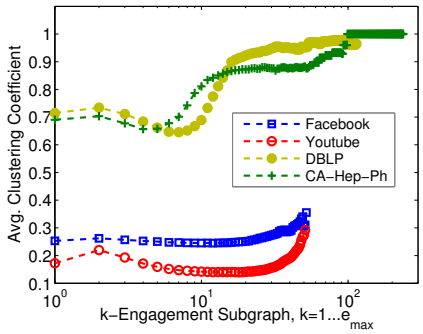
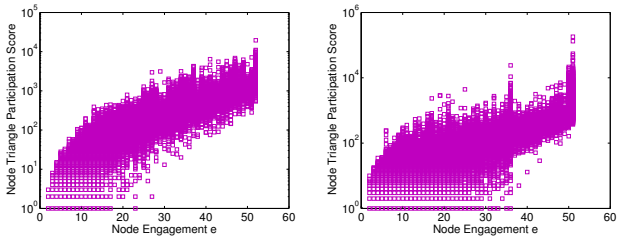


Figure 8: Average clustering coefficient per k-engagement subgraphs $\mathcal{G}_k$. Observe that the clustering coefficient is gradually increasing for larger values of $k = 1, \ldots, e_{\max}$.

Focusing now on the clustering properties of the k-engagement subgraphs, for different values of k ranging from $k = 1, \ldots, e_{\max}$, we examine how the clustering structure (as captured by the clustering coefficient) is affected by the departure of nodes. Figure 8 depicts the average clustering coefficient (CC) of each possible k-engagement subgraph, for four datasets of our collection. As we can observe, the CC increases gradually as we are moving to $\mathcal{G}_k$'s of higher engagement, indicating more cohesive subgraphs with higher degree of interactions among nodes. Actually, this captures the expected behavior of k-engagement subgraphs, in the sense that nodes which belong to higher order $\mathcal{G}_k$ (higher values of k), should

[4]http://en.wikipedia.org/wiki/Clustering_coefficient.



(a) FACEBOOK (b) YOUTUBE

(a) DBLP (b) CA-HEP-PH

Figure 9: Node engagement vs. triangle participation score. Correlation of the node engagement index with the number of triangles that each node participates to.

demonstrate stronger degree of collaboration with their neighbors and thus higher engagement.

Lastly, we consider the clustering properties at node level and we examine how the triangle participation score of each node $i \in V$ (i.e., the number of triangles that each node participates to) is related to the engagement index e_i. Figure 9 presents the correlation of the engagement index e for each node vs. the triangle participation score. It seems that the triangle participation score approximates better the engagement level of a node (compared to the degree), supporting also our intuition that the existence of triangles is vital for the engagement properties.

6. DISENGAGEMENT SOCIAL CONTAGION

In this section, we briefly discuss some potential implications of our observations, regarding the property of engagement. As we presented in Section 4, the engagement of a node is proportional to its core number, and captures its incentive to remain in the graph. As we observed from the experimental results in Section 5, the size distribution of the k-engagement subgraphs is skewed, indicating that a large fraction of nodes typically show a relatively low engagement. Then, based on the size distribution, we were able to characterize the engagement properties of the whole graph. In that case, graphs in which a large portion of their nodes has high engagement level, correspond to more robust graphs in terms of departures. In other words, in such graphs, a relatively high portion of nodes (based on the size of the full graph) do not has incentive to depart.

However, an interesting behavior can possibly occur if we consider a scenario under which nodes can also depart *independently* of their engagement level. In other words, although there is no incentive to depart, nodes decide to drop out possibly due to some external factors. In that case, graphs with high engagement index $\mathcal{E}_G$, will be vulnerable under *targeted* node departures, i.e., when nodes with high engagement value e decide to depart. That is, the "disengagement" cascade that will be formed, will have higher effect in the graph and potentially could cause a relatively large number of nodes to depart as well. In the more general case, according to the skewed size distribution of the k-engagement subgraphs (Fig. 3),

we can say that real graphs tend to be robust under random node departures, but vulnerable to targeted ones. This mainly happens due to the fact that most of the nodes in the graph have relatively low engagement, while a few nodes show high engagement value.

Note that, this type of robustness assessment moves on a similar axis as the well known result about the robustness of real graphs under random/targeted node removals based on their degree. In that case, the seminal result by Albert and Barabási [2] states that, due to the existence of a skewed degree distribution, real graphs are robust against random attacks but vulnerable to targeted ones (in the case of hub nodes). As future work, we plan to experimentally examine how such types of *engagement-based departures* affect the structural characteristics of the graph (e.g., connectivity, diameter).

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied the problem of engagement estimation in a social graph. Based on a game-theoretic model, we proposed several ways to examine the engagement dynamics, at both node level as well as at graph level.

The main contributions of the paper are the following:

- *Problem statement:* We studied the property of engagement in social graphs and we examined how it can be used to model the departure dynamics of the nodes in the graph.
- *Measures of engagement:* We proposed interesting and easy to compute measures for characterizing the engagement dynamics at both node and at graph level.
- *Experiments on real graphs:* We performed several experiments on real-world graphs, observing interesting properties about the engagement dynamics.

As future work, we plan extend our study on more complex types of graphs, such as directed and signed graphs, where the engagement characteristics may behave in a different way. Furthermore, one important point is to further examine this new notion of robustness in real graphs, based on departures of individuals due to their engagement level and not by external attacks or failures.

8. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the constructive comments. Fragkiskos D. Malliaros is a recipient of the Google Europe Fellowship in Graph Mining, and this research is supported in part by this Google Fellowship. Michalis Vazirgiannis is partially supported by the DIGITEO Chair grant LEVETONE in France.

9. REFERENCES

[1] The DBLP Computer Science Bibliography, `http://www.informatik.uni-trier.de/~ley/db/`.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[3] J. I. Alvarez-Hamelin, , L. Dall'Asta, A. Barrat, and A. Vespignani. k-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *NHM*, 3(2):371, 2008.

[4] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, pages 7–15, 2008.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54, 2006.

[6] R. Baeza-Yates and M. Lalmas. User engagement:the network effect matters! In *CIKM*, 2012.

[7] V. Batagelj and M. Zaversnik. An o(m) algorithm for cores decomposition of networks. *CoRR*, 2003.

[8] K. Bhawalkar, J. Kleinberg, K. Lewi, T. Roughgarden, and A. Sharma. Preventing unraveling in social networks: the anchored k-core problem. In *ICALP*, pages 440–451. 2011.

[9] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of internet topology using k-shell decomposition. *PNAS*, 104(27):11150–11154, 2007.

[10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[11] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD*, pages 597–605, 2012.

[12] A. Harkins. Network games with perfect complements. Technical report, University of Warwick, February 2013.

[13] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, Aug 2010.

[14] S. Lattanzi and D. Sivakumar. Affiliation networks. In *STOC*, pages 427–434, 2009.

[15] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1), 2007.

[16] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large welldefined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[17] V. H. Manshadi and R. Johari. Supermodular network games. In *Allerton*, pages 1369–1376, 2009.

[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, pages 29–42, 2007.

[19] B. Pittel, J. Spencer, and N. Wormald. Sudden emergence of a giant k-core in a random graph. *J. Combin. Theory Ser. B*, 67(1):111–151, 1996.

[20] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *ISWC*, pages 351–368, 2003.

[21] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, pages 695–704, 2011.

[22] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

[23] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *PNAS*, 109(16):5962–5966, 2012.

[24] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *WOSN*, pages 37–42, 2009.

[25] S. Wu, A. Das Sarma, A. Fabrikant, S. Lattanzi, and A. Tomkins. Arrival and departure dynamics in social networks. In *WSDM*, pages 233–242, 2013.

[26] Y. Zhang and S. Parthasarathy. Extracting analyzing and visualizing triangle k-core motifs within networks. In *ICDE*, pages 1049–1060, 2012.