Studying the Dissemination of the K-core Influence in Twitter Cascades

Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy

Studying the Dissemination of the K-core Influence in Twitter Cascades

Sarah Elsharkawy[1], Ghada Hassan[2,3], Tarek Nabhan[1], and Mohamed Roushdy[3]

[1] Research and Development Department, ITWORX, EGYPT.
{sarah.elsharkawy,tarek.nabhan}@itworx.com
[2] Faculty of Computer and Information Sciences, The British University in Egypt.
ghada.hassan@bue.edu.eg
[3] Faculty of Computer and Information Sciences, Ain Shams University, Egypt.
mroushdy@cis.asu.edu.eg

Abstract. The k-core of an information graph is a common measure of a node connectedness in diverse applications. The k-core decomposition algorithm categorizes nodes into k-shells based on their connectivity. Previous research claimed that the super-spreaders are those located on the k-core of a social graph and the nodes become of less importance as they get assigned to a k-shell away from the k-core. We aim to evaluate the influence span of the social media super-spreaders, located at the k-core, in terms of the number of k-shells that their influence can reach. We base our methodology on the observation that the k- core size is directly correlated to the graph size under certain conditions. We explain these conditions and then investigate it further on real-life meme cascades extracted from Twitter. We utilize the correlation to assess the effectiveness of the k-core nodes for influence dissemination. The results of the carried-out experiments show that the correlation exists in our studied real-life datasets. A high correlation existed between the k- core size and the sizes of the inner k-shells in all the examined datasets. However, the correlation starts to decrease in the outer k-shells. Further investigations have shown that the k-shells that were less correlated exhibited a higher presence of spam accounts.

1 Introduction

The super-spreaders are the users capable of initiating a viral spread of a piece of information, a meme or an idea. Identifying a set of users as super-spreaders is an indication that they can influence a significantly large number of other users in the cascade. However, it does not indicate whether their influence spread is in the depth or breadth of the social network. For instance, a celebrity account or a news agency account that has thousands of followers could be a super-spreader because they influence many their direct followers, hence, the spread is in the cascade breadth. A depth effect, on the other hand, would signify their ability to influence

the followers of their followers. In this case, they can diffuse viral content to users outside the list of their direct followers.

Previous research [1, 15] has identified the users with largest spreading influence to be the users located in the k-core of the network. The k-core is a maximal subgraph, where all nodes are connected to some number (k) of other nodes in the same subgraph. The k-core decomposition analysis is used to find the k-core of a given graph by iteratively deleting nodes with degree less than k. The degree of a node is the number of edges connected to the node. A k-shell is the subgraph of nodes in the k-core but not in the (k+1)-core. The work done in [9] proved that the most efficient spreaders are those located within the core of the network as identified by the k-shell decomposition analysis. The authors proved their observation by counting the total number of cascaded successors of each node and showing that the number of successors is correlated with the k-value of the k-shell where each user is located. However, their study did not differentiate between the users whose successors span a few k-shells from those whose successors span many k-shells.

In this paper, we propose a measure to estimate how far the influence of the super-spreaders located at the inner-most k-core of the cascade reaches other users located in further k-shells in the cascade. We differentiate between the datasets where the k-core users' influence is confined to the few neighboring k-shells, as opposed to where their influence is disseminating to the shallower k-shells of the cascade. In addition, we investigate whether the presence of spam accounts in the information cascade lessen the overall influence of the k-core users.

The paper is organized as follows: Section 2 gives an overview of research like this work. Section 3 studies the relationship between the k-core size and the graph size on synthetic graphs and discusses the meaning of such relationship in real-life datasets. Section 4 presents the conducted experiments and their results. Finally, Section 5 discusses and concludes this work.

2 Related Work

The virality of memes, as indicated by a large and/or quick growth in size, has been examined from various perspectives. A meme may become viral because it appeals to many [3], but virality of a meme also depends on other factors such as network structure, randomness, adoption patterns of influential users, timing, and many others [12]. Authors of [6] proposed an approach to judge the reliability of the cascade size in social networks by observing the k-core size. They argued that the size of the cascade, on its own, is a misleading indicator of the growth of popularity.

Other lines of research focused on identifying influential spreaders using different measures. Examples of such measures are the number of retweets, the number of followers, the number of mentions, betweenness centrality, and k- core [5, 9]. Authors in [11] found that the best spreaders are consistently located in the k-core across dissimilar social platforms such as Twitter and Facebook.

In [7], the authors evaluated communities based on the k-core concept, as means of evaluating their collaborative nature. In [8], the authors show that k-cores

have an important role in counter-contagions in online social networks. They stated that to start a counter-contagion to an existing contagion, one needs to search for the most influential nodes to start with. k-cores was one of the methods they proposed to identify those influential nodes. Authors of [9] show that the most efficient spreaders are not necessarily the most connected people in the network, but rather are those located within the core of the network as identified by the k-core decomposition analysis.

3 Relationship Between K-core Size and Graph Size

In this section, we discuss the relationship between the inner k-core size and the graph size and its significance in meme cascades. Consider the following two scenarios for two different dynamic meme cascades: in the first cascade, as the k-core size increases, the graph size enlarges as well. And, as the k-core size gets smaller, the graph size decreases too. This synchronization between the k-core size and the graph size signifies that the core users are consistently spreading their influence to the whole community, and this relationship strengthens the criticality of the k-core users. In the second cascade, the k-core size increases and decreases regardless of the growth or shrinkage of the graph size. In this scenario, there is no noticeable relationship between the k-core size and the graph size, which indicates that the core users are not spreading their influence effectively. Therefore, we conclude that the correlation between the k-core size and the graph size is a projection for the real-life influence of the k-core on the rest of the cascade.

3.1 Conditions for the Presence of Correlation

Let G be a simple connected graph and k_d be the k value of the inner-most k-core of G; that is, no core exists at $k = k_d + 1$. Let S(G, k) be the k-core size measured as the number of nodes in the k-core of G. The size of the whole graph can be represented by S(G).

According to [14], the degrees of the nodes in the inner k-core represent an upper bound for the value k_d. Given the degree distribution (DD) of any simple connected graph, there is a percentage (P) of the graph nodes that have a degree larger than or equal to k_d, and a percentage $(100-P)$ of nodes that have a degree less than k_d. Based on the k-core decomposition analysis, all the nodes in the k_d-core must have a degree larger than or equal to k_d, and hence they constitute a portion (P_{core}) of the (P) nodes. The k_d-core size is then calculated as:

$$S(G, k_d) = P_{core} \times P \times S(G) \qquad (1)$$

Let the ratio of the k_d-core size to the graph size be denoted by r. Based on equation 1, a set of graphs $\{G_1, G_2, ..., Gn\}$ having equal percentages $\{P_1 = P_2 = ... = P_n\}$ and equal $\{P_{core_1} = P_{core_2} = ... = P_{core_n}\}$ would consequently have

equal ratios $\{r_1 = r_2 = ... = r_n\}$. Hence, a direct correlation between the k_d-core size and the graph size exists within the given set of graphs.

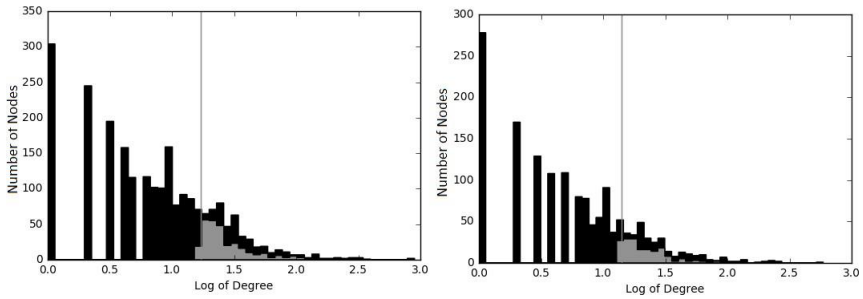

Fig. 1. Histograms representing the degree distributions of two synthetic graphs that follow a power-law DD. The vertical thin grey line is the k_d value. The black bars represent the number of graph nodes and the grey bars represent the number of k_d- core nodes.

3.2 Correlation on Synthetic Scale-Free Power-Law Degree Graphs

In this section, we focus on exploring the correlation on scale-free power-law synthetic graphs as it is the type of graphs found in social networks [2]. We generated thousands of scale-free synthetic graphs and used power-law curve fitting approaches to estimate the exponent of the power-law DD.

We used the Networkx Python package for graph generation and manipulation. Each synthetic graph is generated as follows:

1. A power-law sequence of degrees is generated (using the power law sequence method.)
2. The graph is constructed based on the generated degrees (using the method configuration model).
3. Any self-loops and parallel-edges that may exist on the graph are removed and the largest connected component is extracted.
4. The DD of the graph is fitted to a power-law model, and the exponent of the fitted model is measured using the approach of [4].

Although we have created the graphs based on the power-law sequence of degrees (steps 1 and 2), there is a possibility that the DD of some graphs is not a power-law distribution due to applying step 3. A standard approach to check whether a given distribution follows a certain model is to use the goodness-of-fit test, such as the Kolmogorov-Smirnov (KS) [10], which generates a p-value that quantifies the plausibility of the hypothesis which, in our case, states that the observed data is drawn from a power-law distribution. The KS-test is applied to the graph relative to its fitted power-law model and the graph is disposed if its p-value is larger than 0.05. We ran the experiments using a variety of exponents and other parameters to ensure the variability of graph structures. We clustered the graphs into groups based on the power-law exponent of their fitted model.

We observed that the graphs having the same power-law exponent, also have the same P and P_{core}. For illustration, Figure 1 plots two synthetic graphs that follow a power-law DD, where their k_d values lie on the 78^{th} percentile of the degree range and the number of the k_d-core nodes is approximately 12% of the number of the total graph size.

Table 1 shows the exponents and the k_d-core size to graph size ratio of each group. We notice that, within each group, the k_d-core size to the graph size ratio varies minimally. This indicates a high positive correlation between the k_d-core size and the graph size when the graphs have similar power-law DD.

Table 1. Synthetic graphs grouped based on power-law exponent of node DD.

Exponent	k_d-core size to graph size percent
2.16 ± 0.023	(1.79 ± 0.66) %
2.35 ± 0.037	(1.45 ± 0.64) %
2.65 ± 0.073	(1.5 ± 1.12) %
2.82 ± 0.051	(2.37 ± 3.26) %
3.01 ± 0.051	(11.53 ± 13.58) %
3.2 ± 0.058	(20.88 ± 7.23) %
3.4 ± 0.058	(17.04 ± 3.95) %

3.3 DD Similarity in Real-life Dynamic Cascades

In the previous section, we discussed the correlation on synthetic graphs. Synthetic graphs are constructed using specific rules and hence, they do not model the infinite patterns that occur in real-life datasets. In a real-life meme cascade, users are affected by one or more sources of influence that guide their adoption behavior towards a meme. To model a dynamic meme cascade, we take snap- shots of the cascade as it evolves over time. We found that the set of graphs representing the snapshots of a single meme cascade share similar structural properties, and their DD varies minimally. The mean and standard deviation of the power-law exponents of the snapshots in our datasets are shown in Table 2.

To measure the K_d-core influence dissemination, in terms of k-shells, in a given meme cascade, we propose the following:

1. Taking consecutive snapshots of the meme cascade over periodic time steps.
2. Measuring the correlation between the inner k_d-core sizes and the graph sizes of the snapshots.
3. If the correlation is high: we conclude that the users located in the k_d-core of the given meme cascade are the most influential spreaders. They constitute the dominating source of influence that guides the propagation of the meme cascade.
4. If the correlation is low:

(a) We propose to measure the correlation between the K_d-core size and each of the neighboring k-shells as an attempt to define the portion of the cascade that is being influenced by the K_d-core. The k-shells exhibiting a high correlation are the ones being influenced by the K_d-core.

(b) One of the possible reasons behind the low correlation is the effect of other influential sources on the cascade propagation such as word-of- mouth, media channels, and/or spam. We recommend using a spam detection approach to detect the spam nodes, specifically within the nodes of the k-shells that are less correlated with the K_d -core due to the existence of a good indication that the nodes in these shells are not behaving like the rest of the graph nodes.

Table 2. Datasets Description

Dataset Name		Tsunami	Royal Baby	P1	P2
Time span		10/3/2011 - 10/4/2011	22/6/2013 - 22/7/2013	1/6/2013 - 1/7/2013	1/6/2013 - 1/7/2013
Tweets Count		770,083	137,036	110,782	13,446
Users Count		415,642	130,788	42,760	7,821
Power-law exponent	Mean	3.06	4.63	3.03	3.69
	St. dev.	0.89	0.78	0.68	1.01
Spearman's rho for k_d-core to cascade size		0.767076	0.80355	0.54163	0.47127

4 Experiments and Results

In this section, we present the conducted experiments and their results. In section 4.1, we describe our datasets. Sections 4.2 and 4.3 present our findings.

4.1 Datasets Description

We collected four real-life Twitter datasets using the free Twitter API. The first is Tsunami dataset of tweets discussing the tsunami disaster of 2011 in Japan. The second is the Royal baby dataset which is the set of tweets discussing the birth of the son of Prince William of the UK. Two more datasets P1, P2 were collected that are tweets regarding two concurrent but competing political campaigns that ran during a constrained period. Table 2 shows the time span, the total number of tweets collected and the count of users in each dataset.

For each dataset, the set of information cascades representing the dynamics of the dataset were constructed. Each cascade represents a snapshot of the dataset information cascade with a 24-hour time span between the cascades. For each cascade (snapshot in time), the node degree was fitted, the exponent determined, and the KS-test used to test the goodness-of-fit between the node degree curve and the power-law curve following the procedure described earlier in Section 3.2. Table 2 shows the mean and the standard deviation of the values of the fitted power-law curve exponents for each dataset.

4.2 Correlation on Twitter Datasets

In the first experiment, we measured the correlation between the k_d-core size and graph size of each dataset of snapshots. Table 2 shows the Spearman's correlation coefficients (ρ) between k_d-core size and graph size of each dataset. All the reported coefficients are statistically significant with a P-value < 0.05. We notice that the correlation is high in Tsunami (0.76) and Royal Baby (0.8) datasets, and it drops in the two political datasets, P1 (0.54) and P2 (0.47).

We conducted another experiment in which we correlate the k_d-core size with the size of each k-shell at $k < k_d$, to monitor at which k value the correlation would break. Table 3 shows how the correlation coefficient is high at inner shells and gradually decreases as we move to the outer shells. In Tsunami and Royal Baby datasets the correlation remains relatively high at all shells. However, in P1 and P2 political datasets, the correlation breaks at outermost shells.

4.3 The Spam Effect on the Correlation

The decrease of correlation is a direct indication of the disengagement of the growth or shrink of the k_d-core size and the graph size. One of the possible reasons of this disengagement is the effect of other influencing sources acting on the cascade. Due to the lack of ground-truth information about the other sources that affected our datasets, we are only able to use heuristic approaches that are found in literature to get an estimation of such an external effect on our cascades. After we have constructed all the cascades of our four datasets, we used Truthy's BotOrNot API [13] to check all user accounts in our graphs to determine how likely they are to be spam. We then spotted the location of each of them on the graph and measured the k-shell value at which each of them is located. Results are shown in Figure 2.

The percentage of spam accounts relative to the cascade size is 0.02% in Tsunami, 0.05% in Royal Baby, 0.11% in P1 and 0.1% in P2. Figure 2 shows the amount of spam detected at each k-shell of the four datasets. We observe that no spam is found at inner k-shells while it increases at outer shells. The results match the values found in Table 3 which show how the correlation is significantly high at inner k-shells and decreases at outer shells.

Table 3. Spearman's correlation coefficients between k_d-core size and each outer k-shell size measured on all snapshots of each dataset.

(k) of Shell	Tsunami	Royal Baby	P1	P2
	Shell to 13-core	Shell to 29-core	Shell to 29-core	Shell to 28-core
1-shell	0.654	0.783	0.538	0.550
2-shell	0.750	0.789	0.545	0.532
3-shell	0.751	0.702	0.555	0.537
4-shell	0.721	0.782	0.571	0.656
5-shell	0.762	0.784	0.592	0.701
6-shell	0.815	0.784	0.655	0.849
7-shell	0.822	0.785	0.671	0.859

8-shell	0.819	0.821	0.728	-
9-shell	0.823	0.829	0.732	-
10-shell	0.877	0.831	0.739	-
11-shell	0.886	0.836	0.745	-
12-shell	0.890	0.835	0.752	-
13-shell	-	0.839	0.755	-
14-shell	-	0.842	0.757	-
15-shell	-	0.845	0.761	-
16-shell	-	0.845	0.767	-
17-shell	-	0.845	0.812	-
18-shell	-	0.851	0.861	-
19-shell	-	0.854	0.888	-
20-shell	-	0.855	0.895	-
21-shell	-	0.854	-	-
22-shell	-	0.855	-	-
23-shell	-	0.857	-	-
24-shell	-	0.860	-	-
25-shell	-	0.861	-	-
26-shell	-	0.866	-	-
27-shell	-	0.866	-	-
28-shell	-	0.872	-	-

From this experiment, we uncovered three main findings: first, the number of spam accounts found in each dataset is inversely proportional to the correlation between the K_d-core and the graph sizes of the snapshots of a given dataset. Where Tsunami has $\rho = 0.76$ and spam ratio of 0.02%, Royal baby has $\rho = 0.8$ and spam ratio of 0.05%, P1's $\rho = 0.54$ and spam ratio of 0.11 and finally P2 has $\rho = 0.47$ and spam ratio of 0.1. Second, the number of spam accounts in each k-shell is inversely proportional to the correlation between the k_d-core and k-shell sizes of the dataset as seen in Table 3 and Figure 2. The third is that the number of spam accounts increases at outer k-shells and almost vanishes at the inner K_d-core.

5 Discussion and Conclusion

In this paper, we studied the relationship between the inner k-core size and the graph size and we tackled the question of how far the influence of the most influential spreaders located at the inner k-core reach users located in the outer k-shells.

We presented a novel approach to estimate the influence reach of the users located at the k-core inferred from the observed correlation between the k-core size and the graph size. We demonstrated that a correlation between the inner k-core size and the cascade size exist in sets of synthetic graphs under some constraints, and we identify such constraints.

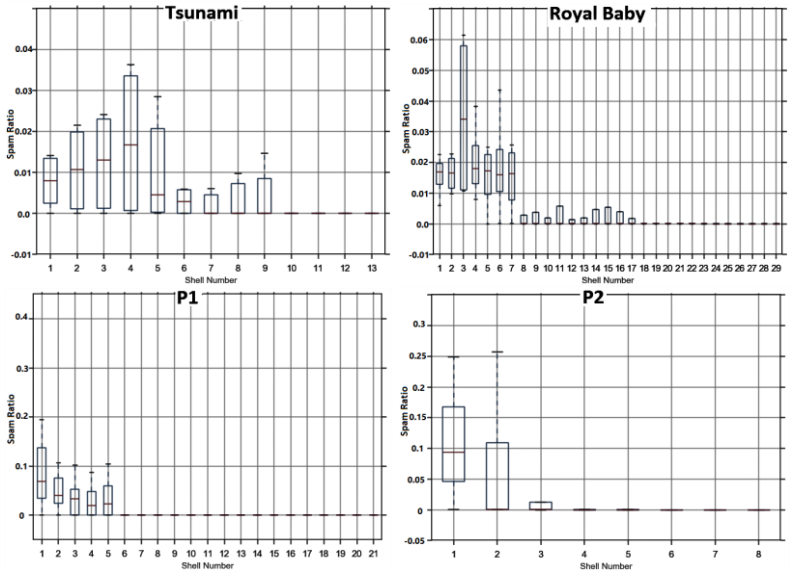

Fig. 2. Box plot representing the probability distribution of the percentage ratio between the spam count in each k-shell to the shell size. The x-axis represents the (k) value of the k-shell, and the y-axis represents the percentage ratio of the number of spam accounts in a given k-shell to the number of nodes in the same k-shell.

We presented results of a case study on four real-life Twitter datasets. The datasets represented snapshots of the meme propagation graph every 24 hours. We found that the correlation between the k-core and the neighboring k-shells decreases gradually towards the outer shells. We also found that in political datasets the correlation drops significantly at outer shells due to external influencing factors such as spam, while in the other two datasets the correlation remained consistent. Using Truthy BotOrNot API, we identified the spam accounts in our datasets. We found that the number of spam increases at outer shells and that the number of spam accounts in a shell is inversely proportional to the correlation of the size of that shell to the k-core size.

We conclude that the influence propagation of the super-spreaders located at the k-core of a meme cascade varies significantly from one dataset to another. In the datasets which exhibit other sources of influence such as spam accounts, the super-spreaders are notably unable to disseminate their influence to the shallower k-shells of the cascade. These datasets are signified by the low correlation between the inner k-core size and the graph size of snapshots captured over periodic time steps of the lifetime of the cascade. For those datasets, having low correlation, marketers need to pay more attention to other types of super-spreaders such as those having high betweenness centrality. Identifying the ideal super-spreaders that should be targeted in this case is still an area of investigation.

Another important conclusion is that the datasets that have a low correlation between its inner k-core size and graph size usually unveil a large number of spam accounts. Hence, the proposed correlation measure could be used for an early warning for spam. Moreover, we observed that the outer k-shells have a higher tendency

for the presence of spam. This phenomenon could be used to speed up the spam detection algorithms by searching in the outer k-shells first.

References

1. M. A. Al-garadi, K. D. Varathan, and S. D. Ravana. Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. Physica A: Statistical Mechanics and its Applications, 468(C):278–288, 2017.
2. A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.
3. J. Berger and K. L. Milkman. What makes online content viral? Journal of Marketing Research, 49(2):192–205, 2012.
4. A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. SIAM review, 51(4):661–703, 2009.
5. S. Dorogovtsev, A. Goltsev, and J. Mendes. k-core organization of complex networks. Physical Review Letters, 96:4, 2006.
6. S. Elsharkawy, G. Hassan, T. Nabhan, and M. Roushdy. On the reliability of cascade size as a virality measure. Proceedings of the European Conference on Electrical Engineering and Computer Science (EECS), 2017.
7. C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, pages 87–93, 2011.
8. A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: a survey. ACM SIGMOD Record, 42(2):17–28, 2013.
9. M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identification of influential spreaders in complex networks. Nature Physics, 6:888–893, 2010.
10. F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253):68–78, 1951.
11. S. Pei, L. Muchnik, J. S. A. Jr., Z. Zheng, and H. A. Makse. Searching for superspreaders of information in real-world social media. Scientific Reports, 4:5547, 2014.
12. H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, pages 365–374, New York, NY, USA, 2013. ACM.
13. J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, pages 249–252, 2011.
14. K. Shin, T. Eliassi-Rad, and C. Faloutsos. Corescope: Graph mining using k-core analysis - patterns, anomalies and algorithms. ICDM, pages 469–478, 2016.
15. F. Zhang, Y. Zhang, L. Qin, W. Zhang, and X. Lin. Finding critical users for social network engagement: The collapsed k-core problem. pages 245–251, 2017.