Uncovering the Nucleus of Social Networks

Braulio Dumba, Zhi-Li Zhang
University of Minnesota, Twin Cities, MN, USA
braulio,zhzhang@cs.umn.edu

ABSTRACT

Many social network studies have focused on identifying communities through clustering or partitioning a large social network into smaller parts. While community structure is important in social network analysis, relatively little attention has been paid to the problem of "core structure" analysis in many social networks. Intuitively, one may expect that many social networks possess some sort of a "core" which holds various parts of the network (or constituent "communities") together. We believe that it is just as important to uncover and extract the "core" structure – referred to as the "nucleus" in this paper – of a social network as to identify its community structure. In this paper, we propose a scalable and effective procedure to uncover the "nucleus" of social networks by building upon and generalizing ideas from the existing k-shell decomposition approach. We employ our approach to uncover the nucleus in several example communication, collaboration, interaction, location-based and online social networks. Our methodology is very scalable and can also be applied to massive networks (hundreds million nodes and billion edges).

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Human-centered computing** → *Social network analysis*; • **Theory of computation** → Shortest paths;

KEYWORDS

Social Network; K-Shell Decomposition; Network Core

ACM Reference Format:
Braulio Dumba, Zhi-Li Zhang. 2018. Uncovering the Nucleus of Social Networks. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3201064.3201075

1 INTRODUCTION

Networks are often abstractly modelled as a graph where vertices represent entities and edges capture the relations (e.g., connections) or interactions between them. In the context of (online) social networks, community identification has received a lot of attention. A community is often considered to be a subset of vertices that are densely connected internally but sparsely connected to the rest of the network [9, 30, 35–37]. The majority of studies on identifying communities structures in social networks have relied on clustering techniques, namely, by partitioning the underlying network/social graph into *disjoint* (sometimes *overlapping*) communities. For example, Newman proposes a measure of betweenness – modularity [36, 37] – for identifying disjoint communities in a social network. Andersen et al [9] design a local graph partitioning algorithm to indentify community structures. This algorithm is based on personalized PageRank vectors. Ahn et al [6] introduce a novel perspective for discovering hierarchical community structures by categorizing links only. To obtain an optimal partition and to find communities at multiple levels, an information-theoretic framework is proposed by the authors in [38, 40]. Several studies use link and content information for uncovering meaningful communities in networks [22, 50].

Although existing studies of community structure have been very successful, most have not considered the existence of "core structure" in many networks. Intuitively, one expects that many social networks possess some sort of "core" as part of their meso-scale structure, which holds various parts of the network (or constituent "communities") together. We believe that it is just as important to uncover and extract the "core" structure – referred to as the "nucleus" – of a social network as identify its community structure [39, 49]: unlike "ordinary" constituent communities, the "core" structure plays a crucial role in the formation and evolution of a social network, to which other (constituent) "communities" are attached. Chung and Lu [18] show that power-law random graphs almost surely contain a core "subgraph" when the exponent β in the power-law degree distribution is such that $\beta \in (2, 3)$. This theoretical result suggests that many real-world social networks likely posess some sort of cohesive core structure.

One of the most popular notion of network core is given by the *k-shell decomposition* method [15]. This classical graph decomposition technique decomposes a network into hierarchically ordered layers from the periphery to the core. This method has also be extended to weighted graphs [24, 48] and dynamic networks [32]. The k-shell decomposition method has often been used as a visualization tool for studying the core structure of massive complex networks such as the Internet [15]. In addition, it has been used to identify influential spreaders in a network [23, 28].

When applying the standard k-shell decomposition to uncover the core of several example social networks (see § 2), we find that the resulting "innermost" structure is unlikely to represent the "core" of these networks. For example, this "innermost" structure may contain the maximum clique of a network but which lies rather at its periphery, or it is simply a single vertex in a dense graph. This appears to the effect of the (iterative) degree-based pruning process of k-shell decomposition, where despite at some point we reach the vicinity of the core, the k-shell decomposition continues further, which then destroys the "core" structure of the network (see § 3

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201075

for more illustration). This raises the following important question: *When should we stop the k-shell decomposition pruning process in order to preserve the core graph G_C of a network?*

In an attempt to address this question, we develop an effective procedure to uncover the *nucleus* structure of a social network by building upon and generalizing ideas from the existing k-shell decomposition [15] approach, as follows. Firstly, we propose a new metric, the *dependence value*, that measures the location importance of a node in a network. Intuitively, the dependence of node v captures the number of nodes recursively dependent of v that have been removed in earlier steps of the k-shell decomposition method. Secondly, we derive a new measure called *nucleon-index* (NI) that captures the extend to which a subgraph is a densely intra-connected and topological central core. This index can be used with a wide variety of functions to transition between core and peripheral nodes (e.g., dependence value, closeness [41] and betweenness [41] centralities, etc). Using these metrics, we therefore modify the standard k-shell decomposition method to stop the process earlier, in order to extract a meaningful "core" for social networks (see § 4). For a Facebook [4, 29] friendship network composed of 63,731 nodes and 817,035 edges, this process yields a dense "core" subgraph G_C with approximately 285 nodes and 9,616 edges. Given a dense core subgraph G_C, we investigate the importance of this substructure for the network by analysing the following metrics (see § 5): i) the distance between a node v to the core subgraph G_C; ii) the ratio of the distance between nodes u and v to their respective distance to G_C and iii) lastly, the impact of removing G_C in the structure of the network G ($G_C \subset G$).

We discuss implications and related work in § 6 and § 7. Section 8 concludes the paper. We summarize the major contributions of our paper as follows:

- We propose two *new* metrics: i) the *dependence value*, that measures the location importance of a node in the network; ii) the *nucleon-index* (NI) that captures the extend to which a subgraph is a densely intra-connected and topological central core . Using these metrics, we therefore modify the standard k-shell decomposition method to stop the process earlier, in order to extract a meaningful "core" for social networks.
- We apply our approach to uncover the core structure in example communication, collaboration, interaction, location-based and online social networks. Our methodology is very scalable and can also be applied to uncover the core structure of massive networks (hundreds million nodes and billion edges).

2 DATASETS

This section presents a summary of the datasets that we use for our analysis:

Autonomous systems graph: This dataset is an undirected graph of the AS peering information inferred from Oregon route-views between March 31 and May 26, 2001 [2], and its main features are summarized on Table 1.

Social networks graphs: This dataset is a collection of 9 undirected graphs of communication, collaboration, interaction, location-based and online social networks [1–5, 11, 25, 29, 46](see Table 1 for a summary of the main features):

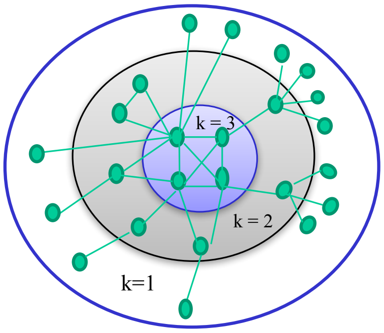
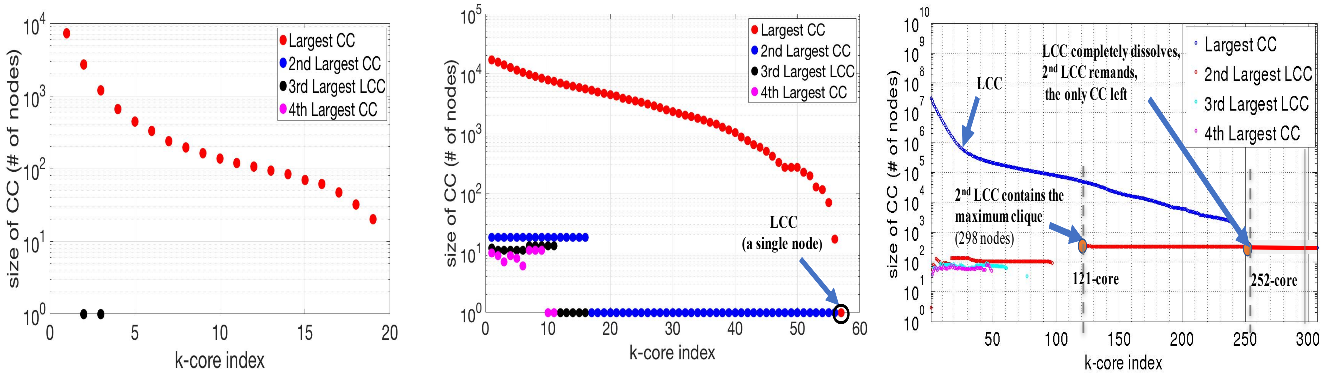


Figure 1: A schematic representation of a network under k-shell decomposition: the network can be viewed as the union of shell 1 up to $k_{max} = 3$ (network core).

Table 1: Main characteristics of the social networks and AS graphs: d - node degree; % LCC - percentage size of the largest connected component of the original network

ID	# nodes	# edges	max(d)	% LCC
arenas-jazz	198	2,742	100	1.00
dnc-corecipient	906	20,858	368	0.94
arenas-pgp	10,680	24,316	205	1.00
Oregon-1	11,174	23,409	2,389	1.00
ca-HepPh	12,008	118,521	491	0.93
ca-AstroPh	18,722	198,110	504	0.95
ca-CondMat	23,133	93,497	280	0.92
email-Enron	36,692	183,831	1,383	0.92
loc-brightkite	58,228	214,078	1,134	0.97
Facebook	63,731	817,035	1,098	0.99

- *ca-AstroPh, ca-HepPh, ca-CondMat*: collaboration networks between authors for papers submitted to Astro Physics, High Energy Physics (Phenomenology category) and Condense Matter Physics – a graph contains an undirected edge (i, j), if author i co-authored a paper with author j.
- *arenas-jazz*: collaboration network between jazz musicians – the graph contains an undirected edge (i, j), if two musicians have played together in a band.
- *email-Enron*: email communication network – the graph contains an undirected edge (i, j), if address i sent at least one email to address j.
- *arenas-pgp*: interaction network of users of the Pretty Good Privacy (PGP) algorithm.
- *dnc-corecipient*: online contact network for people having received the same email in the 2016 Democratic National Committee email leak – the graph contains an undirected edge (i, j), if two persons received the same email.
- *Facebook*: an undirected subgraph of the friendship network for the users in Facebook.
- *loc-brightkite*: an undirected graph for the friendship network for the users from loc-brightkite location-based online social network.



(a) Oregon-1

(b) ca-AstroPh

(c) Google+

Figure 2: The size of the largest as well as those of the 2nd, 3rd and 4th largest connected components in the k-core subgraphs

3 K-SHELL NETWORK CORE

K-shell decomposition [15] is one of the most popular and scalable method to investigate and visualize the core-periphery structure in complex networks. This method assigns to each node an integer representing its coreness location according to successive layers or shells in the network. It works as follows: a) first, remove all nodes in the network with degree 1 (and their respective edges) – these nodes are assigned to the 1-shell; b) more generally, at step $k = 2, \ldots$, remove all nodes in the remaining network with degree k or less (and their respective edges) – these nodes are assigned to the k-shell; and c) the process stops when all nodes are removed at the last step. Small values of k define the periphery of the network and the *innermost network core* corresponds to the highest shell index (k_{max}) – see Fig. 1. (Note that this is distinct from *k-core decomposition*[1] defined in the literature [7, 8]).

In the k-shell decomposition process, at each step k, the remaining subgraph is referred to as "k-core" (C_k). The k-core subgraph is the union of all shells with indices larger or equal to k or it is the maximal induced subgraph $C_k \subseteq G$ such that if $v \in C_k$, then node v must have at least $k + 1$ neighbors that belong to C_{k-1} and $deg^k(v) > 0$ (we use $deg(v)$ to denote the degree of v in the network and $deg^k(v)$ to denote the degree of v in C_k). Similarly, k-shell (S_k) can be defined as the subgraph induced by the set of nodes with $d^{k-1}(v) \leq k$ and if $v \in S_k \rightarrow deg^k(v) = 0$.

Clearly, for a node to belong to the k-core (thus $shell(v) \geq k$), it must have at least degree k, i.e., $deg(v) \geq k$. However, $deg(v) \geq k$ is not sufficient to guarantee it to belong to the k-core. For example, a node v with only neighbors of degree 1 (i.e., v is the root of a star structure) belongs to the 2-shell, i.e., $shell(v) = 2$, no matter how high its degree is. On the other hand, it is easy to see that if a node v is part of a clique of k nodes, then $shell(v) \geq k$. However, a node v does not need to be part of a k-clique to have $shell(v) \geq k$. Consider a tree T of n nodes (the sparsest graph with n nodes). We can in fact provide a complete characterization of nodes in T to have $shell(v) \geq k$ in a recursive manner: for v to have $shell(v) \geq k$, it must have at least k-neighbors u's with $shell(u) \geq k - 1$ – this characterization also applies to a general graph. We see that in the case of a tree, nodes with higher k-shell indices must lie more at the "core" (i.e., the increasingly "denser" part) of the tree. For a general graph, however, a node with a high k-shell index may not lie at the "core" of the graph: it can be part of a large clique that is "isolated" on a periphery of a massive graph. In such a case, the large clique will break off from the "core" of the network (e.g., as represented by the largest connected component remaining in the k-core) in the early stage of the k-shell decomposition process.

This method has been successfully used as a visualization tool for studying and uncovering the core structure of networks such as the Internet AS graph [15]. We apply it to the *Oregon-1* AS dataset. Fig. 2(a) shows the size of the largest as well as those of the 2nd, 3rd and 4th largest connected components in the k-core graph. We observe that the largest connected component decreases smoothly as k varies from 1 to 20. At $k_{max} = 20$, we are left with a very dense core subgraph composed of 20 nodes and 164 edges – the network nucleus. This result shows that for the AS graph, nodes with the highest k-shell indices indeed lie at the "core" (i.e., the increasingly "denser" part) of the graph. However, our experiments reveal that applying the k-shell decomposition for other types of graphs, especially social graphs, may not yield the same results. There are two possible reasons:

First, for some graphs the k_{max}-shell seems to contain some "residual" portions of the nucleus of a graph or simply a singleton node. For example, Fig. 2(b) shows the k-core graph for the 4 largest connected components in the *ca-AstroPh* dataset. We see that at k_{max}=57, we are left with just a single node in the k-core graph, which is unlikely to be the complete inner-core of the graph.

Second, in other graphs the k_{max}-shell does not appear to lie at the "core" of the graph: it could be part of a large community structure (e.g. a maximum clique) that is "isolated" on a periphery of a graph. To illustrate this, we apply the k-shell decomposition method to a *Google+* reciprocal network[2] obtained from a previous

[1] Which simply removes all nodes with degree less than k in a graph.

[2] A network composed with only bi-directional edges, extracted from a directed social graph. A reciprocal network can be viewed as the stable "skeleton" network of a directed social network that holds it together and encodes its main topological characteristics [20]. For more on the reciprocal network of Google+ the reader is referred to [19, 20].

study [19, 20] - it consists of more than 40 million nodes and ≈ 400 million edges. Figure 2(c) shows the size of the largest as well as those of the 2nd, 3rd and 4th largest connected components in the k-core, as k varies from 1 to 308. We note that at step $k = 121$, a small subgraph containing the maximum clique (of size 290) breaks off from the largest connected component which desolves after $k = 253$, whereas this subgraph containing the maximum clique persists after $k = 252$ and becomes the largest component; at $k_{max} = 308$, we are left with this maximum clique plus 10 additional nodes that are connected to the maximum clique. Closer inspection of the nodes in the maximum clique reveals that its users belong to a single institution in Taiwan, forming a close-knit community where each user follows everyone else – which is unlikely to be the network core of Google+.

From these results, we see that directly applying the standard k-shell decomposition to some graphs (especially, social networks) produces an "innermost" structure that does not represent "core" of these networks. This is due to the fact that at a certain k-index, we reach the vicinity of the core; but going far beyond this index would destroy the core structure of the network.

4 NODE DEPENCENCE VALUES AND NETWORK CORE

In order to extract a meaningful "core" for a general graph $G = (V, E)$ (e.g., social networks), we therefore modify the standard k-shell decomposition method to stop the process earlier. To achieve this, we propose a new metric that provides important information about the structural function of each node in the graph (we label it as "dependence" value) at each k-step. Then, we present a new measure called *nucleon-index* (NI) that captures the extend to which a subgraph is a densely intra-connected and topological central core – it can be used with a wide variety of functions to transition between core and peripheral nodes (e.g., dependence value, closeness and betweenness centralities, etc).

4.1 Node Depencence Values

The *dependence* value of node v at step k is defined as follows: for $v \in V$, $dep^0(v, \beta) = 0$ and for $k = 1, \ldots, c(v)$,

$$dep^k(v,\beta) := dep^{k-1}(v,\beta) + \delta^k(v) + \beta \times \Sigma_{u \in N^k(v)}[dep^{k-1}(u,\beta)] \tag{1}$$

where β is a control parameter, $0 \le \beta \le 1$; $N^k(v)$ is the set of neighbors of node v that are removed at step k, and $\delta^k(v) = |N^k(v)|$. The dependency of node v is recursively defined by measuring the number of nodes u (the h-hop neighbors of v, $h = 1, \ldots, k$) that are removed in earlier steps up to $k = c(v)$ –the *coreness* of node v (and for $k \ge c(v)$, by convention, we define $dep^k(v, \beta) = dep^{c(v)}(v, \beta)$).

Intuitively, $dep^k(v, \beta)$ captures the number of nodes recursively dependent on v that have been removed in earlier steps up to k. With $\beta = 0$, we note that $dep^k(v, \beta)$ captures the number of v's neighbors removed at each step up to k, and for $k \ge c(v)$, $dep^k(v, \beta) = \sum_k \delta^k(v) = deg(v)$, the degree of node v. With $\beta > 0$, $dep^k(v, \beta)$ captures not simply the dependence of its neighbors, but that of its neighbors' neighbors, and so forth. However, the number of nodes u removed at each step up to k does not influence the dependence value of the node v uniformly. Their contribution is weighted by the parameter β in eq.(1). The parameter β quantifies the contribution of node u to the total dependence value of node v. More precisely, at the kth-step, we multiply the number of h-step removed neighbors of v by β^{h-1} (see the proof in the appendix). Thus, the further a node u is to node v, the less it will contribute to the total dependence value of node v. Hence, a node v having more nodes u with high dependence values in its vicinity will also have a high dependence value, creating the *dependency propagation* effect. Therefore, we posit that the network core should contain only nodes with very high dependence because the $dep^k(v, \beta)$ values of any $v \in V$ grows as k increases (more nodes are removed as we move from the periphery of the graph to its core). In the next section, we use the dependence value of node v as a measure of its coreness.

Table 2: *Arenas – jazz*: **peak nucleon-indices (*NI*) and their respective k_C-indices (set *SK*) and β values**

β	max(NI)	k_C
0.0	0.011019	26
0.1	0.006561	25
0.2	0.006125	24
0.3	0.006841	24
0.4	0.007256	24
0.5	0.007500	24
0.6	0.007818	25
0.7	0.008545	25
0.8	0.009222	25
0.9	0.009849	25
1.0	0.010433	25

4.2 Nucleon Index and Network Nucleus

To derive a meaningful "core" structure in social networks, we postulate that the *nucleus* of a network $G(V, E)$ is an induced subgraph G_C having the following properties:

1. Subgraph $G_C(V_C, E_C)$ is *connected* and composed of a collection of nodes in G with *dense* aggregate centralities by some measure.
2. The set V_C is fundamental for the *structural properties* of the network, e.g., in terms of connecting nodes via short paths through the network.
3. G_C is the minimal subgraph with these properties.

To find a subgraph G_C with the above properties, we consider an appropriately defined "decomposition" process (e.g., the k-shell decomposition) which yields a (filtration) sequence of (sub)graphs $\{G_k\}$'s of G: $G_0 := G \supset G_1 \supset \cdots \supset G_K = \emptyset$. Given a node centrality measure $\theta(i)$, $i \in V$, we define the *nucleon-index* (NI) to capture the extent to which a subgraph constitutes a "densely connected", topological central core in this sequence:

$$NI(G_k, \theta(i)) := \frac{V_k}{V_{k-1}} \times \frac{E_k}{V_k \times (V_k - 1)} \times \{\frac{1}{V_k} \times \sum_{i \in G_k} \theta(i)\} \tag{2}$$

where by abuse of notation, we use E_k to denote the number of edges between nodes in G_k and V_k the number of nodes in G_k (and $|V_K| = 0$). The second term in eq.(2) measure the density of G_k and the last term the average centrality of G_k. Ideally, if G_k