Covertness Centrality in Networks

Michael Ovelgönne
UMIACS
University of Maryland
College Park, MD 20742
mov@umiacs.umd.edu

Chanhyun Kang, Anshul Sawant
Department of Computer Science
University of Maryland
College Park, MD 20742
chanhyun@cs.umd.edu, asawant@cs.umd.edu

VS Subrahmanian
UMIACS & Department of Computer Science
University of Maryland
College Park, MD 20742
vs@cs.umd.edu

***Abstract*—It has been known for some time that in terror networks, money laundering networks, and criminal networks, "important" players want to stay "off" the radar. They need sufficient centrality (according to traditional measures) to be well connected with the rest of their network, but need to blend in with the crowd. In this paper, we propose the concept of covertness centrality (CC). The covertness centrality of a vertex v consists of two parts: how "common" v is w.r.t. a set $\mathcal{C}$ of centrality measures, and how well v can "communicate" with a user-specified set of vertices. The more "common" v is, the more able it is to stay hidden in a crowd. Given $\mathcal{C}$, we first propose some general properties we would like a common-ness measure to satisfy. We then develop a probabilistic model of common-ness that a vertex has w.r.t. $\mathcal{C}$ (specifying, intuitively, how many other vertices are like it according to all centrality measures in $\mathcal{C}$). Covertness centrality of vertex v is then defined as a linear combination of common-ness and the ability of v to communicate with a user-specified set of other vertices. We develop a prototype implementation of CC and report on experiments we have conducted with it on several real-world data sets.**

I. Introduction

Social networking research has focused extensively on the problem of identifying important vertices in a network, taking into account the structural properties of the network. As a consequence, important concepts such as degree centrality, between-ness centrality, closeness centrality, and eigenvector centrality (see [1]) have been proposed (as well as many others that we do not list here). Work in this arena dates back to the beginning of the last century.

However, less work has focused on characterizing vertices that can communicate well with a specified set of vertices, while having sufficiently low centrality that they do not "stick out". There are certainly several important applications where users want to have low centrality according to classical centrality measures, but still retain the ability to communicate with a given set of vertices. For example: (i) Terrorists or criminals communicating through online social networks may need to communicate with a certain set of vertices, i.e. their terror or criminal network, while appear "common", i.e. looking similar to many other vertices in the network so that they cannot be easily distinguished from other vertices. (ii) A booming industry is in online marketing on platforms like Twitter where companies offer to "insert" certain types of messaging into the network for a client. In such cases, the marketers want to get the message out without "standing" out as spreaders of a marketing message. We call actors who wish to satisfy (i) and (ii) *covert actors*, i.e. actors who want to communicate well with certain other actors, but who wish to stay "below the radar."

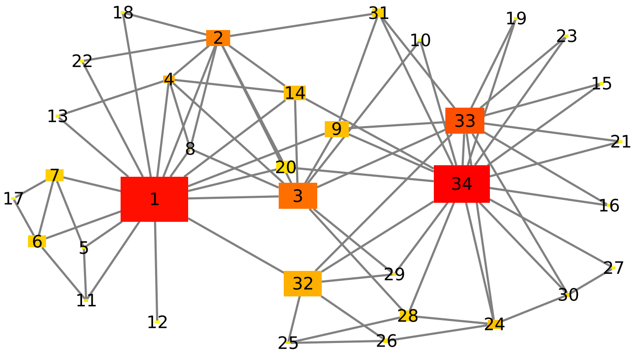
In this paper, we will define a new kind of centrality measure, denoted as *covertness centrality* (CC for short), that captures these intuitions. Unlike classical centrality measures which start solely with a graph as input, CC assumes that we are given as input, both a graph *and* a set $\mathcal{C}$ of classical centrality measures. Intuitively, think of $\mathcal{C}$ as being a set of centrality measures that a covert actor thinks an adversary might use to identify key players in a network.

We split the definition of CC into two parts - how "common" is a vertex in a network, given $\mathcal{C}$? And how well can he communicate with a given set of nodes in the network according to the centrality measures in $\mathcal{C}$? The answer to the second question, of course, has been extensively studied below.

However, we note that this tension between the opposing goals of common-ness and strong ability to communicate has been studied before. In pioneering work, Lindelauf et al. [2] analyzed optimal network structures with respect to information and secrecy. Depending on different scenarios and measures for information and secrecy, a wheel graph with a center that is connected to all vertices of a ring-like structure was shown to be optimal. Gutfraind [3] came to similar results when analyzing cascade resilience of terrorist networks. But is such an organizational structure actually favorable? The center of a wheel structure has very unique properties (e.g. a high degree and high betweenness). Having an observable property that is unique or rare decreases the effort necessary to be identified assuming the respective property is known by the prosecutors.

Figure 1 shows how common-ness can be measured in the well-known Karate club network [4]. Vertices 13, 15–19, 21–23 and 27 have identical degree and betweenness centrality scores. Moreover, vertices 5, 8, 11, 13, 26, 29, 30 and 31 have degree and between-ness scores that are also very similar to this group. This means that every vertex in this set of 18 vertices can hardly be distinguished from the others w.r.t. their network position. However, we note that vertices 1 and 34 have very similar degree centrality scores but the betweenness score of vertex 1 is much higher than the score of vertex 34. An interesting vertex is vertex 32. This vertex is similar to vertices 3 and 33 w.r.t. betweenness centrality and similar to

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.156



id	dc	bc	id	dc	bc
1	10.3	29.3	18	1.3	0
2	5.8	3.6	19	1.3	0
3	6.4	9.6	20	1.9	2.2
4	3.8	0.8	21	1.3	0
5	1.9	0.0	22	1.3	0
6	2.6	2.0	23	1.3	0
7	2.6	2.0	24	3.2	1.2
8	2.6	0.0	25	1.9	0.1
9	3.2	3.7	26	1.9	0.3
10	1.3	0.1	27	1.3	0
11	1.9	0.0	28	2.6	1.5
12	0.6	0	29	1.9	0.1
13	1.3	0	30	2.6	0.2
14	3.2	3.1	31	2.6	1.0
15	1.3	0	32	3.8	9.2
16	1.3	0	33	7.7	9.7
17	1.3	0	34	10.9	20.3

Fig. 1. Zachary's well-known karate club network [4]. The size of the vertices is proportional to the betweenness centrality (higher score, larger vertex). The vertex color shows the degree centrality (color from yellow (low) to red (high)). The table shows the degree centrality (dc) and betweenness centrality (bc) scores of the vertices.

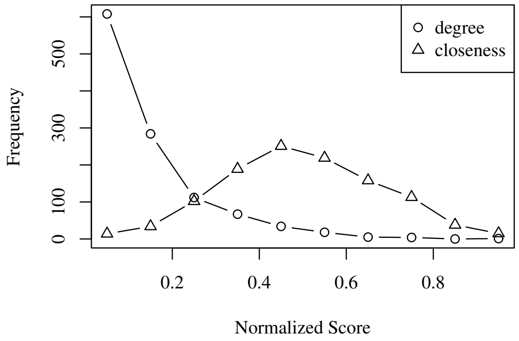


Fig. 2. Histograms of the degree and closeness centrality scores (normalized to the interval [0,1]) for the URV dataset (see Tab. I).

the vertices 4, 9, 14, 24 w.r.t. degree centrality. However, this vertex is not similar to any other vertex w.r.t. both measures.

Measuring which network positions are "not common" and which ones are "common" is not as straightforward as it may seem. In most social networks there are a lot more vertices with a low degree than vertices with a high degree, as most social networks are known to be scale-free. But low values are not always dominating. Closeness centrality scores are usually normally distributed (or follow at least another bell shaped distribution) as shown in Fig. 2 [1]. So it takes more to measure how well a vertex hides in the crowd than looked at how low its centrality scores are.

The rest of this paper is organized as follows: In Section II, we will provide a brief overview of other peoples' work on centrality measures and other vertex scoring functions that can be used to describe the position of an actor in a network. We will discuss requirements that a definition of "common-ness" should satisfy as well as possible ways to define covertness centrality in Section III. Next, in Section V, we will evaluate the proposed covertness centrality measures and algorithms to compute them. Related work will be discussed in Section VI, before we finally draw a conclusion of the work in this paper and provide directions for future work in Section VII.

II. Classical Centrality Measures and Other Vertex Scoring Functions

The position in a network and the ability to efficiently communicate are usually measured with centrality measures. Mathematically, centrality measures are functions $f : V \rightarrow \mathbb{R}$ that assign centrality scores to the vertices of a network. Conceptually they come from the intuition that a network has a center and a periphery. While being at the center means having an important or influential position, peripheral positions come with an opposite status.

Some of the most widely used centrality measures are degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. They respectively measure the number of edges adjacent to a vertex, the average length of the shortest-paths to all other vertices in the graph, and the fraction of shortest-paths between all pairs of vertices that pass through a vertex. A good introduction to these measures provides [1]. Many variations of the four aforementioned measures exists, e.g. variants of betweenness centrality that take only paths up to specific length into account or normalized with respect to path length [5]. Beside the four most popular centrality measures, various other centrality measures have been proposed. For example, the n-path centrality [6] counts the number of paths up to length n that start at a vertex. Gómez et al. [7] took a different perspective from the preceding authors and based their analysis on game-theory. They defined a family of measures of power in networks based on the Shapley value. Lindelauf [8] presented a specific centrality measure based on the Shapley value that he used to analyze covert networks. For a comprehensive classification and comparison of centrality measures see [9].

But there are also other vertex scoring functions than centrality measures that describe the structural position of a vertex. An example is the clustering coefficient [10] that

[1] These distributions for the respective centrality measures hold true for all of our evaluation networks listed in Tab. I

measures the fraction of existing to all possible edges between the neighbors of vertex. *Throughout this paper, we use the term centrality measure to include any function $f : V \to \mathbb{R}$ whose intent is to compute the influence or importance of a node as exemplified by such measures.*

III. Covertness Centrality Definition

Our notion of covertness centrality is a combination of commonness and communication potential. Commonness means hiding in a crowd of equal or similar actors and communication potential is the ability to efficiently communicate in order to achieve the objective of the group. In this section, we will discuss how to construct a covertness centrality measure from this two components.

A. Commonness

As introduced, commonness should measure how well an actor hides in a crowd of similar actors. Let $CM(a)$ denote the (as yet undefined) commonness of an actor a. Intuitively, we want $CM(a) > CM(b)$ if and only if knowledge of properties of the structural position of actor a reveals less information about a than knowledge of properties of the structural position of actor b reveals on b. This means, that if the size of the crowd actor a is hiding in is larger than the crowd actor b is hiding in, a's covertness centrality has to be higher.

In terms of conditional probability we can describe what we want to measure as follows. If we want to search for an actor a and we know that this actor has the property I, what is the probability that a randomly picked actor with property I is actor a? In this work, we are focusing solely on network data. That means, no other information on actors are available than the edges connecting them. Therefore, the possible information on a person are properties of his or her position in the network. Moreover, we assume that this information is encoded via a user-specified set $\mathcal{C}$ of centrality measures.

Before we can give a formal definition of covertness centrality, we need to introduce some definitions. Let $G = (V, E)$ be a graph where V denotes a set of vertices and $E \subset V \times V$ and set of edges connecting pairs of vertices. Furthermore, let $\mathcal{C} = (C_1, C_2, \ldots, C_k)$ be an ordered set of centrality measures on graph G where C_i is a function from V to $\mathbb{R}$.

We can regard an actor w as similar to an actor v with respect to a centrality measure C_i, when $C_i(w)$ and $C_i(v)$ are "similar". We define what values of C_i can be regarded as similar based on the variance of the values of the centrality measure. Let σ_i be the standard deviation of centrality scores $C_i(v)$ where $v \in V$. Then, the k-dimensional interval of centrality values similar to the one of vertex v (v's similarity neighborhood) is

$$\begin{aligned} I_v = &[C_1(v) - \alpha\sigma_1, C_1(v) + \alpha\sigma_1] \times \\ &[C_2(v) - \alpha\sigma_2, C_2(v) + \alpha\sigma_2] \times \\ &\ldots \times \\ &[C_k(v) - \alpha\sigma_k, C_k(v) + \alpha\sigma_k] . \end{aligned}$$

where $\alpha \geq 0$ is a parameter that defines the range of similar values. A user can define what α should be.

Furthermore, let $X(v) = (C_1(v), C_2(v), \ldots, C_k(v))$ be the vector of centrality measures for vertex v and let $\sigma(v) = (\sigma_1(v), \sigma_2(v), \ldots, \sigma_k(v))$ be the vector of the standard deviations for all centrality measures in $\mathcal{C}$.

As we have a finite set of actors, the distinct centrality vectors that are similar to a vertex v w.r.t. $\mathcal{C}$ are

$$D_{\mathcal{C}}(v) = \{X | C_i(w) = X_i, i \in \{1..k\}, w \in V \wedge X \in I_v\}.$$

The probability that a randomly chosen vertex has the centrality vector X is

$$\begin{aligned} f_{\mathcal{C}}(X) = &P(C_1(v) = X_1 \wedge \ldots \wedge C_k(v) = X_k) \\ = &\frac{|\{v | C_1(v) = X_1 \ldots \wedge C_k(v) = X_k\}|}{|V|} . \end{aligned}$$

However, in our context, we are not interested in the probability that a randomly chosen vertex has a specific centrality vector but in the probability that some other randomly chosen vertex has the same centrality vector X as a given vertex. So we define

$$g_{\mathcal{C}}(X) = \max \left\{ \frac{|\{v | C_1(v) = X_1 \ldots \wedge C_k(v) = X_k\}| - 1}{|V| - 1}, 0 \right\} .$$

Given these definitions, we can define the commonness $CM(v)$ of a vertex $v \in V$ with respect to a non-empty set $\mathcal{C}$ of base centrality measures in various ways. However, it is desirable (though not strictly necessary) that all specific definitions of CM satisfy certain properties.

Property 1. *Optimal Hiding: If all vertices are equal w.r.t. all centrality measures, the vertices are indistinguishable and so the hiding is optimal. In this case the commonness has to be* 1 *for all vertices.*

$$\forall v, w \in V, C \in \mathcal{C} \quad C(v) = C(w) \Rightarrow \forall u \in V : CM(u) = 1.$$

Optimal hiding gives the upper limit for commonness as there can be no better hiding than in the crowd that consists of all other actors.

Property 2. *No hiding: Given a definition of similarity, if a vertex is not similar to any other vertex w.r.t. any centrality measure, the commonness of this vertex should be 0.*

$$\forall w \neq v \in V, C \in \mathcal{C} \quad C(v) \nsim C(w) \Rightarrow CM(v) = 0$$

This property defines the lower bound for commonness and simply means that if an actor cannot hide at all, s/he is maximally exposed and should get the lowest possible commonness score.

Property 3. *If one measure in the set of base centrality measures assigns the same score to all vertices, removing this measure from the set should have no effect on the commonness scores.*

$$\exists x \forall v \in V \quad \overline{C}(v) = x \Rightarrow CM_{\mathcal{C}}(v) = CM_{\mathcal{C} \setminus \overline{C}}(v)$$

This axiom basically says that if a centrality measure $C \in \mathcal{C}$ is not able to distinguish between the vertices (i.e. it assigns the exact same centrality score to all vertices), than it should have no influence on the commonness score.

Given this desired properties, we can define specific commonness measures and see how well they align with these properties.

Definition 1 (Common-ness Measure CM_1). *We can define commonness as the sum of the squared distances separately for each dimension:*

$$CM_1(\mathcal{C}, v) = 1 - \frac{\sum_{C \in \mathcal{C}} \left(1 - \sum_{X \in D_{\{C\}}(v)} g_{\{C\}}(X)\right)^2}{k} \quad (1)$$

Definition 2 (Common-ness Measure CM_2). *Another way to define commonness is the fraction of all vertices that are similar to a vertex v in all considered dimensions:*

$$CM_2(\mathcal{C}, v) = \sum_{X \in D_{\mathcal{C}}(v)} g_{\mathcal{C}}(X). \quad (2)$$

If we think of the deviation from no informational value (that is when all actors have the same centrality value) as an error, we can define commonness as the sum of the squared error over all dimensions as in Definition 1. However, if the centrality measures are uncorrelated, measuring the averageness of an actor independently for each measure can lead to undesired results. For example, an actor might appear to be totally average w.r.t. all measures considered independently, but if the sets of actors he is similar to are large, but do not overlap for the different centrality measure, there is no crowd he can hide in. To account for this, in Definition 2 the joint probability distribution of all centrality measures is used.

We now investigate whether the two definitions have the desired properties. We start with CM_1.

Proposition 1. *CM_1 satisfies properties (1) and (2) but not (3).*

Proof Sketch. It follows immediately for the definition that CM_1 satisfies Property 1. If for a centrality measure all vertices have the same value x, then for all centrality measures C there is a X so that for all vertices v $D_{\{C\}}(v) = X$ and $f(X) = 1$. It follows immediately $CM_1(v) = 1 - (1/1) = 0$. Property 2 is satisfied as well. If for a measure C and a vertex v for all other vertices $w \in V$ $C(w) \notin I_v$ (this is the definition of similarity $\sim$ of CM_1), then $D_{\{C\}}(v)$ is empty. So, the sum over all X in $D_{\{C\}}$ is 0, the whole fraction becomes $k/k = 1$ and $CM(v) = 0$. Finally, lets have a look at Property 3 for CM_1. Let us assume we have two centrality measures C_1 and C_2, whereof C_2 is non-distinguishing. For C_1 the term $\sum_{X \in D_{\{C\}}(v)} g_{\{C\}}(X)$ will have some value $a < 1$ for every v while the term get 0 for C_2 (compare discussion of Property 2). So $CM_1(\{C_1\}, v) = 1 - a/1$ while $CM_1(\{C_1, C_2\}, v) = 1 - (a + 1)/2$. Since $a < 1$ follows $CM_1(\{C_1\}, v) \neq CM_1(\{C_1, C_2\}, v)$ and CM_1 does not satisfy Property 3.

Proposition 2. *CM_2 satisfies all 3 properties (1)–(3).*

Proof Sketch. The proof that CM_2 satisfies Property 1 is similar to that for CM_1. If for some set of centrality measures $\mathcal{C}$ all vertices have the identical centrality vector X, for every $v \in V$ the only similar vector $D_C(v)$ is X. As all vectors are identical $g(x) = 1$ and so $CM_2(\mathcal{C}, v) = 1$ for every v. Likewise, if $D_{\{C\}}(v)$ is the empty set when there are not other vertices similar to a vertex v, i.e. there is no $w \neq v \in V$ with $C(w) \in I_v$. From $D_{\{C\}}(v) = \emptyset$ follows $CM_2(\cdot, v) = 0$ and so CM_2 satisfies Property 2. To proof that Property 3 is satisfied goes as follows. Let C_n be a measure that assign the same score s to every vertex. If we add C_n to $\mathcal{C}$ we only extend the centrality vectors and intervals by one dimension. As every vertex has the same centrality score of C_n, for every vector $Y = (y_1, \ldots, y_k)$ in $D_{\mathcal{C}}(v)$ there is a corresponding vector $Y_n = (y_1, \ldots, y_k, s)$ in $D_{\mathcal{C} \cup C_n}(v)$. As $X = (x_1, \ldots, x_k)$ we be extended to $X_n = (x_1, \ldots, x_k, s)$ it follows that $g_{\mathcal{C}}(X) = g_{\mathcal{C} \cup C_n}(X_n)$ holds because of the correspondence between X and X_n.

Thus, CM_2 satisfies all three properties we want a commonness measure to have, while CM_1 satisfies only the first two desired properties. As a consequence, CM_2 seems to be the epistemologically better choice.

B. Communication Potential

The communication potential should reflect the ability to communicate and cooperate to achieve a common objective. This vague statement makes no point about which communication and cooperation options are important to achieve the common goal. If only in-group connections are important for achieving the group's objective, we define the the communication potential based on a centrality measure D and the group $\hat{V} \subset V$. Let $\hat{G} = (\hat{V}, \hat{E})$ be the induced subgraph of G given by the group $\hat{V}$. Then the communication potential is $CP_1(v) = D_{\hat{G}}(v)$, i.e. the centrality score of v on the graph $\hat{G}$ that is "of interest". In a criminal or terrorist network application, $\hat{V}$ might consist of vertices that an investigative agency has already uncovered - people they know are suspicious for one reason or another — and the agency wants to "uncover" the rest of the covert network. However, if the ability to communicate with people outside the group is important as well, the entire graph G can be used to measure the communication potential: $CP_2(v) = D_G(v)$.

Calculating CP_1 requires knowledge of the group that a vertex is trying to communicate with — a factor that can be problematic in the real world. Therefore, we can only calculate and compare CP_1 and so the covertness centrality of the members of a group to each other. CP_1 is undefined for actors that are not group members. So a covertness centrality measure based on CP_1 is suitable for explaining organizational structures of terror/criminal groups when $\hat{V}$ is known through other processes. In contrast, a covertness centrality score based on CP_2 is defined for every actor in a network. It can be used to answer generic questions such as "Where would criminals try to hide?"

C. Covertness Centrality

In Section III-A and III-B we discussed definitions of the two components of covertness centrality: commonness and communication potential. There is no unique "best way" to combine these two measure into one — any combination method would reflect a different judgment on the importance of common-ness and communication. Let us assume that the communication potential CP is normalized in a reasonable way to the interval $[0, 1]$ like CM is. A possible family of functions is

$$CC(v, \tau, \lambda) = \begin{cases} 0 & , CM(v) < \tau \\ \lambda CM(v) + (1 - \lambda) CP(v) & , CM(v) \geq \tau \end{cases} \tag{3}$$

For $\tau = 0$, CC is a classic trade-off between the two requirements. Additionally requiring a minimum level τ of commonness would reflect an understanding that the communication potential is irrelevant as long as the threat of being identified is too high. For the case of criminal networks, the actual assessment of the trade-off is influenced by the type of illegal activity a groups pursues. Morselli et al. [11] discuss that criminal for-profit organizations put more emphasis on efficiency, while ideological groups like terrorists are primarily concerned with security.

Let us return to the graph depicted in Fig. 1 and see which vertices have the highest covertness centrality in that example. If we use degree and betweenness as the base centrality measures (i.e. as the set $\mathcal{C}$), degree centrality as the measure for communication potential and weight commonness and communication potential equally ($\lambda = 0.5$, $\tau = 0$), vertices 9, 14 and 20 have the highest covertness centrality scores (0.89, 0.89 and 0.86) for CM_1 (with a difference of 0.1 to the next vertices). For CM_2 the vertices 9 and 14 have the highest values scores (0.77 and 0.74). Vertex 20 has a slightly lower score (0.69) and other vertices follow with a distance of 0.05. It is easy to related this result to the figure: the vertices 9, 14 and 20 are still decently hidden but have position nearer to the center of the network.

IV. CC Computation using Sampling Methods

Calculating the exact commonness of a vertex requires calculating all base centralities for all vertices. While the computational effort to calculate some centrality measures is low, e.g. degree centrality, other centrality measures have a high runtime complexity, e.g. closeness and between-ness centrality. Calculating the exact commonness for large graphs is a very time-consuming or practical impossible task. However, if we are not interested in the covertness centrality of all vertices but only for some, we can speed up the calculation.

We could estimate the probability $g_{\mathcal{C}}(X)$ that a random vertex has a given centrality score from a sample of vertices $S \subset V$. That means, we replace $g_{\mathcal{C}}(X)$ with

$$\tilde{g}_{\mathcal{C}}(X) = \min \left\{ \frac{|\{v \in S | C_1(v) = X_1 \ldots \wedge C_k(v) = X_k\}| - 1}{|S| - 1} \right\}. \tag{4}$$

If we expect that the centrality scores follow a power-law distribution (as they e.g. usually do for degree and betweenness centrality [12]), this sampling strategy might be problematic. We can expect random sampling – and hence $\tilde{g}_{\mathcal{C}}(X)$ – to tend to underestimate the size of the neighborhood for vertices in the long tails of power-law distributions when the sample size is low. In a similar but less significant way, centrality measures involving other distributions (e.g. closeness centrality distributions which are usually bell-shaped) can be affected when large parts of the score range have very low probabilities.

Let $\hat{f}_{\mathcal{C}}$ be the estimated joint probability distribution for the set of centrality measure $\mathcal{C}$. The continuous equivalent of CM_1 (Definition 1) is

$$\widehat{CM}_1(\mathcal{C}, v) = 1 - \frac{\sum_{i=1}^{k} \left(1 - \int_{C_i(v)-\alpha\sigma_i}^{C_i(v)+\alpha\sigma_i} \hat{f}_{\{C_i\}}(X) dX\right)^2}{k}. \tag{5}$$

and the equivalent of CM_2 (Definition 2) is

$$\widehat{CM}_2(\mathcal{C}, v) = \int_{C_1(v)-\alpha\sigma_1}^{C_1(v)+\alpha\sigma_1} \ldots \int_{C_k(v)-\alpha\sigma_k}^{C_k(v)+\alpha\sigma_k} \hat{f}_{\mathcal{C}}(X) dX_k \ldots dX_1 \tag{6}$$

The equations are identical to their discrete counterparts except for the calculation of the estimated number of vertices in the interval around v. The equivalence goes directly back to the derivation of the integral as the approximation of a stepwise function with infinite steps.

As we can assume that many centrality measures have a power-law distribution and are at least to some degree correlated to the degree centrality, an improved sampling technique might also improve the accuracy of heuristic algorithms. Instead of simple random sampling, systematic sampling [13] based on the degree of the vertices seems to be a good idea. Simple random sampling means we randomly select vertices of V. For systematic sampling, we order all vertices by degree. From a population of n vertices, we then select k vertices by taking every n/k-th vertex starting from a randomly select start vertex among the first n/k-th vertices.

To summarize, the computation of commonness is based on the size of the similarity neighborhood of a vertex, which can be computed in several ways:

1) Exact computation: Empirical distribution of all vertices in V
2) Heuristic computation based on simple sampling:
    a) Empirical distribution[2]
    b) Estimated distribution[3]
3) Heuristic computation based on systematic sampling:
    a) Empirical distribution
    b) Estimated distribution

Algorithm 1 shows the pseudo-code to compute covertness centrality for a set of vertices S based on the empirical distributions of a set $\mathcal{C}$ of base centrality measures. The function

[2] i.e. we compute the exact number of similar vertices based on the empirical distribution of the centrality scores

[3] i.e. we compute the expected number of similar vertices based on the probability density of the estimated distribution of the centrality scores

$sample$ could be either a simple random or systematic sampling or could return all vertices V if we want to compute exact scores. The communication potential measure D can be any centrality measure, $\hat{G}$ is an arbitrary subgraph of G (including $\hat{G} = G$) to compute D on, CM is a commonness function like those given in Eq. 1, 2, 5 and 6. The covertness function CC combines commonness and communication potential as in Eq. 3.

Though the Covertness Centrality Algorithm (Algorithm 1) can take any distribution as input, we note that the computation of commonness based on estimated distributions requires a method to detect the type of distribution of centrality measure as well as the parameters of the distribution — due to space constraints, we do not discuss 2(b) and 3(b) above.

```
Algorithm 1: Covertness Centrality
input : Graph $G = (V, E)$, set of vertices $S$, set of
        centrality measures $\mathcal{C}$, communication potential
        measure $D$, sampling function $sampling$,
        commonness function $CM$, covertness function
        $CC$, subgraph $\hat{G}$
output: covertness centrality scores list<Float>
        $covertness$
▼ GetBaseCentralityScores
    for $v \in sample(V) \cup S$ do
        for $C \in \mathcal{C}$ do
            $scores[v][C] \leftarrow C(G, v)$
▼ GetCovertnessCentralityScores
    for $v \in S$ do
        $cm \leftarrow CM(v, scores)$
        $cp \leftarrow D(v, \hat{G})$
        $covertness[v] \leftarrow CC(cm, cp)$
    return $covertness$
```

A. Algorithmic Complexity

The covertness of actors is especially of interest for the analysis of large networks as only then is there potentially a large crowd to hide in. Therefore, the computational complexity of algorithms to compute covertness centrality is important. To compute the exact commonness of a vertex, the base centrality score of all vertices for all centrality measures in $\mathcal{C}$ must be computed and the most expensive one will determine the compound complexity. Eigenvector and shortest-path based centrality measures have at least a quadratic runtime. However, for most measures, approximation algorithms have been developed. For a discussion of centrality measure computation see [14].

Let n denote the number of vertices in a graph. Calculating the exact commonness has a complexity equal to the sum of the base centrality measures plus the time to determine the neighborhood of a vertex. If all the centrality vectors are stored in a range tree whose creation takes $O(n(\log n)^{k-1})$, retrieving the similarity neighborhood of a vertex takes $O((\log n)^k + t)$ where t is the number of retrieved neighbors [15]. To calculate the covertness centrality of all vertices only for all distinct centrality vectors, a neighborhood search is required. In most cases, the dominating factor in the complexity of common-ness will originate from the calculation of the base centrality measures, e.g. when the set of base centralities includes a shortest-path based measure.

For some base centrality measures, sampling can speed up the calculation of common-ness. Instead of calculating the empirical distribution of the centrality of all vertices, the empirical distribution of a random sample can be used as shown (see Equation 4). This is beneficial for all centrality measures where the individual centrality scores can be calculated independently for each vertex. In the case of betweenness centrality, all source shortest paths have to be computed for getting the score of one vertex and so individual score computation provides almost no benefit. For eigenvector centrality, computing the score of only one vertex is impossible as the scores of the vertices are interdependent. If all scores have to be calculated, sampling would only decrease the complexity of the similarity neighborhood search. For closeness centrality however, calculating the scores of only a sample of vertices can dramatically decrease the runtime because in this case the single source shortest path problem has to be solved for each and every vertex.

As discussed in Section IV the simple strategy of using the empirical distribution of a random sample will probably tend to underestimate the size of the neighborhood for vertices in the long tails of power-law distributions. Estimating the parameter of a distribution based on the sample is likely to provide higher accuracy for the same sample size. A second advantage would be that instead of having to determine the similarity neighborhood of each vertex, we just need to calculate the integral in Equation 5 or 6. For a graph with n vertices, a sample size of e.g. $\log n$ would decrease the effort for getting the required closeness centrality scores from $O(n^3)$ to $O(n^2 \log n)$. We will evaluate required sample sizes in Section V. The required sample size depends on the sampling technique and the desired accuracy.

Systematic sampling instead of simple random sampling would require an additional $O(n \log n)$ to sort the vertices. For example, determining the closeness centrality score of one vertex takes $O(n^2)$ time — so the advanced sampling technique would pay off if the same accuracy is achieved with a minimal smaller sample size.

V. Evaluation

In this evaluation, we analyze the properties of the covertness centrality measures as well as the algorithms to calculate them. For this evaluation, we use the testbed of real-world networks listed in Table I.

A. Measures

First, we study the results of the two different commonness definitions. We have already discussed the fact that CM_1 might have a problem when the different centralities in $\mathcal{C}$ are

TABLE I

REAL-WORLD EVALUATION DATASETS. THE YOUTUBE NETWORKS ARE SNOWBALL SAMPLES FROM THE ORIGINAL DATASET. URV IS THE ABBREVIATION FOR UNIVERSITAT ROVIRA I VIRGILI.

Network	#Vertices	#Edges	Type	Reference
URV	1133	10902	e-mail	[16]
Youtube 40k	39998	85793	friendship	[17]
Youtube 60k	59998	151481	friendship	[17]

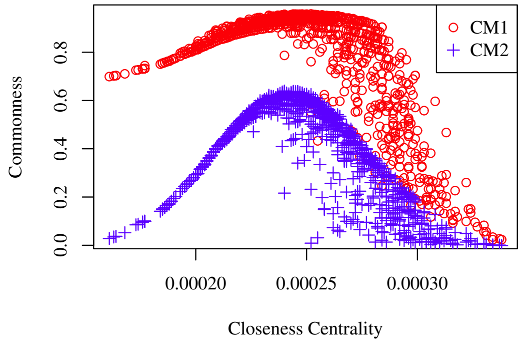


Fig. 3. Scatter plot of the commonness scores according to CM_1 and CM_2 (based on the four base centrality measures degree, closeness, betweenness and eigenvector centrality and the URV dataset) in relation to closeness centrality.

not correlated because vertices that are very exposed w.r.t. one measure might still end up with high commonness scores. This is what we see in Fig. 3. Vertices with low closeness centrality scores have low CM_2 scores because of the low probability density of closeness in that area. However, these vertices have high CM_1 scores as they have many vertices in their similarity neighborhood for other base measures. That compensates for their rare closeness values.

Commonness is strongly negatively correlated to the base centrality measure degree, closeness, eigenvector and betweenness centrality. In Fig. 4 we see for $\lambda = 1$ (i.e. covertness is measured by common-ness only) that the distribution density increases for higher common-ness scores. For lower values of λ, i.e. for an increasing contribution of communication potential to covertness centrality, the distribution of covertness gets closer to that one of the communication potential measure. For $\lambda = 0.25$ the covertness centrality distribution is already very similar to the closeness centrality distribution (compare Fig. 2).

B. Compute Time

We implemented CM_1 and CM_2 in Python and evaluated the performance on a standard desktop machine. The scores of the base centralities were read from a file to evaluate the performance of the covertness computation only. The runtime scales linearly with the number of vertices with compute times of 0.1, 2, and 3 seconds respectively for our three test networks.

Beside the exact commonness algorithm we also studied the heuristic computation of the sizes of the similarity neighborhoods based on sampling. Fig. 5 shows for both sampling methods the Kendall τ rank correlation [18] of the commonness scores to the exact scores (with no sampling) is very high — well over 0.95. To achieve the same accuracy than simple random sample, systematic sampling requires a significantly lower sample size. More detailed results are shown in Table II and Tab. III.

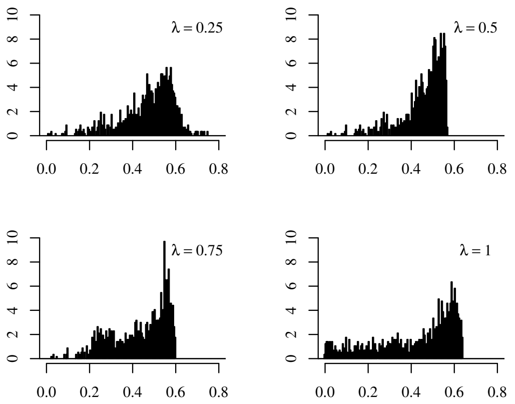


Fig. 4. Distribution of CC scores depending on different λ values when CM_2 (based on degree, betweenness, closeness and eigenvector centrality) is used as the measure of commonness and closeness centrality is used to measure CP and computed base on the whole graph.

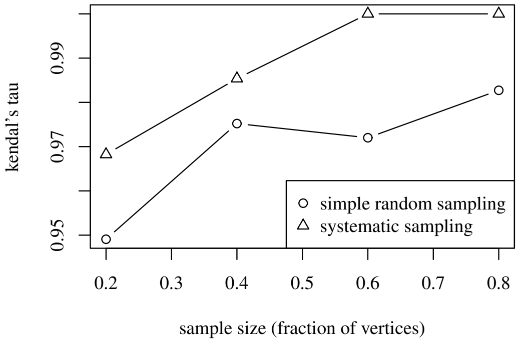


Fig. 5. Comparison of the rank correlation between the exact algorithm and the two sampling algorithm with different sampling methods for the URV dataset.

TABLE II

ACCURACY OF SAMPLING FOR DIFFERENT SAMPLING METHODS AND NETWORKS FOR CM_1 MEASURED WITH KENDAL'S τ.

Network	Simple			Systematic		
	20%	40%	60%	20%	40%	60%
URV	0.965	0.981	0.981	0.966	0.985	1.000
YouTube 40k	0.975	0.992	0.997	0.979	0.996	1.000
YouTube 60k	0.981	0.995	0.997	0.990	0.996	1.000

TABLE III

ACCURACY OF SAMPLING FOR DIFFERENT SAMPLING METHODS AND NETWORKS FOR CM_2 MEASURED WITH KENDAL'S τ.

Network	Simple			Systematic		
	20%	40%	60%	20%	40%	60%
URV	0.880	0.929	0.940	0.876	0.965	0.976
YouTube 40k	0.983	0.985	0.998	0.993	0.992	0.996
YouTube 60k	0.992	0.989	0.996	0.995	1.000	0.998

VI. Related work

The literature on different types of dark, hidden or covert networks of criminal or terrorist groups is extensive. However, most existing literature deals with the network among the members of a covert group only. In contrast, the situation we deal with in this paper is a large network (e.g. phone, e-mail) in which a small fraction of actors constitute a covert subnetwork in a larger (mostly innocent) population.

Our work is inspired by previous work on the conflict between the efficiency of communication in a terrorist group and the attempt to maintain secrecy. Lindelauf et al. [2] analyzed the trade-off between secrecy and information. Secrecy is defined by the exposure probability and the link detection probability. The exposure probability is the probability that a group member gets identified. Link detection probability is the probability that a covert actor is identified by following a link of an already identified covert actor. The link detection probability refers to what Gutfraind [3] denote as cascade resilience. In their case, a link is established if the resulting communication efficiency outweighs the costs of potentially being identified by following that link. In contrast, we see the costs of establishing a link with respect to the exposure an actor receives through it. So our approach is different, as for us covertness does not mean having the least possible number of connections but having connections that appear unsuspicious.

There has been a small amount of work on detecting covert cells within networks. In such a case, a group of individuals tries to stay hidden within a large network. However, most past works on covert cell detection do not study the relationship between centrality measures and covert cell structure. For instance, [19] discusses groups that try to camouflage communications - but the idea that an adversary trying to uncover such covert groups may look at centrality measures is not considered. In addition, there is work on detecting anomalous positions in a network - however, an anomaly may occur not because someone is trying to hide while maintaining communications with cell members but because an individual is in a location in the network that has skewed statistical properties. Conversely, covertness is not anomalous - an individual trying to stay "off the radar" is making an explicit attempt to look "normal" within the network and our notion of common-ness represents an explicit attempt by a "bad guy" to not look anomalous. Thus, the two concepts are very different.

VII. Conclusion

This paper discusses the problem of constructing a measure that reflects covertness of a vertex that combines elements of "common-ness" as well as communication efficiency in a social network is. Being covert in a network means hiding in a crowd of similar actors. Our work is motivated by attempt to analyze how malicious actors in a network behave if they are aware that the network is being monitored via a set $\mathcal{C}$ of centrality measures. Previous work on covertness focused on how actors as parts of small covert networks optimize their position within the network only. We extended those work by the broader view on large networks where covert network are embedded in.

We presented several ways to construct a covertness centrality measure as well as ways to compute the scores of such measures. Given that the base centrality scores are known covertness centrality can be computed efficiently. If the base scores are unknown, we showed that with suitable sampling techniques it is not necessary to determine the exact similarity neighborhoods. We developed an algorithm to compute covertness centrality of nodes using different types of sampling methods, showing that these sampling methods are highly correlated to methods without sampling (Kendall τ coefficient of over 0.95) and that they can be computed relatively efficiently with small sample sizes.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambrige: Cambridge University Press, 1994.

[2] R. Lindelauf, P. Borm, and H. Hamers, "The influence of secrecy on the communication structure of covert networks," *Social Networks*, vol. 31, no. 2, pp. 126–137, 2009.

[3] A. Gutfraind, "Optimizing topological cascade resilience based on the structure of terrorist networks," *PLoS ONE*, vol. 5, no. 11, 2010.

[4] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[5] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008.

[6] D. S. Sade, "Sociometrics of macaca mulatta iii: n-path centrality in grooming networks," *Social Networks*, vol. 11, no. 3, pp. 273–292, 1989.

[7] D. Gómez, E. González-Arangena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada, "Centrality and power in social networks: a game theoretic approach," *Mathematical Social Sciences*, vol. 46, no. 1, pp. 27–54, 2003.

[8] R. Lindelauf, "Design and analysis of covert networks, affiliations and projects," Ph.D. dissertation, Tilburg School of Economics and management, 2011.

[9] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466–484, 2006.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[11] C. Morselli, C. Gigure, and K. Petit, "The efficiency/security trade-off in criminal networks," *Social Networks*, vol. 29, no. 1, pp. 143 – 153, 2007.

[12] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, "Classification of scale-free networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 583–12 588, 2002.

[13] W. G. Madow and L. H. Madow, "On the theory of systematic sampling, i," *The Annals of Mathematical Statistics*, vol. 15, no. 1, pp. 1–24, 1944.

[14] R. Jacob, D. Koschtzki, K. Lehmann, L. Peeters, and D. Tenfelde-Podehl, "Algorithms for centrality indices," in *Network Analysis*, U. Brandes and T. Erlebach, Eds. Springer Berlin / Heidelberg, 2005, vol. 3418, pp. 62–82.

[15] M. Berg, O. Cheong, M. Kreveld, and M. Overmars, *Computational Geometry*. Springer Berlin Heidelberg, 2008, ch. 5 – Orthogonal Range Searching, pp. 95–120.

[16] R. Guimerà, L. Danon, A. Dìaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, p. 065103, Dec 2003.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[18] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, 1938.

[19] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Wallace, "Discovering hidden groups in communication networks," in *Intelligence and Security Informatics*, H. Chen, R. Moore, D. Zeng, and J. Leavitt, Eds. Springer Berlin / Heidelberg, 2004, vol. 3073, pp. 378–389.