Identification of Essential Proteins Using Induced Stars in Protein-Protein Interaction Networks

Chrysafis Vogiatzis*

Department of Industrial & Manufacturing Engineering,
North Dakota State University,
Fargo, ND, USA

Mustafa Can Camur†

Department of Industrial & Manufacturing Engineering,
North Dakota State University,
Fargo, ND, USA

Abstract

In this work, we propose a novel centrality metric (referred to as *star centrality*), which incorporates information from the closed neighborhood of a node, rather than strictly from the node itself, when calculating its topological importance. More specifically, we focus on *degree* centrality and show that in the complex protein-protein interaction networks it is a naive metric that can lead to misclassifying protein importance. For the extension of degree centrality when considering stars, we derive its computational complexity, provide a mathematical formulation, and propose two approximation algorithms. We portray

*E-mail: `chrysafis.vogiatzis@ndsu.edu`; Corresponding author

†E-mail: `mustafa.camur@ndsu.edu`

the success of the new metric in protein-protein interaction networks when predicting protein essentiality in several organisms, including the well-studied *Saccharomyces Cerevisiae*, *Helicobacter Pyloris*, and *Homo Sapiens*, where star centrality is significantly better at detecting essential proteins when compared to nodal centrality metrics. We also analyze the average and worst case performance of the two approximation algorithms in practice, and show that they are viable options for computing star centrality in very large-scale protein-protein interaction networks, such as the human proteome, where exact methodologies are bound to be time consuming.

Keywords. centrality; protein-protein interaction networks; complex network analysis

1 Introduction

Complex network theory has been a driving force for numerous applications in recent years. Many disciplines have observed a paradigm shift after incorporating complex network analysis and graph theoretic tools to interpret their results. One such field is computational biology; the use of complex network analysis has recently enabled researchers with novel tools to detect protein complexes (Li *et al.* 2010, Mitra *et al.* 2013), analyze protein essentiality (Ren *et al.* 2011), and predict protein functionality (Typas and Sourjik 2015), among others. Moreover, the fact that large-scale databases are readily available (we indicatively mention the curated works by Franceschini *et al.* (2013), Szklarczyk *et al.* (2014), Pagel *et al.* (2005), and Salwinski *et al.* (2004)) has led to significant scientific interest in the field in an attempt to tackle the computational and biological challenges inherent to the study of protein-protein interaction networks.

The challenge we focus on in this work can be summarized as follows: does there exist a network topology metric that captures the importance of a single protein in the grand scheme of the proteome? This is not a novel question, as it has attracted numerous researchers and has led to the investigation of various metrics, ranging from graph modularity (Narayanan *et al.* 2011) to centrality (Hahn and Kern 2005). Being able to use such objective metrics for studying the proteome is of importance, as it can lead us to the detection of informal groups in the interaction network (Pereira-Leal *et al.* 2004).

With the term "detection of *informal groups*" we mean the detection of *unbiased* clusters of proteins, based solely on their interactions and topological structure. This would enable us with objective methods of measuring protein importance in the proteome without relying on scientific and experimental biases. In general, topological importance (also broadly referred to as *centrality*) is a well-studied topic in complex networks, including protein-protein interaction networks. In our work, though, we propose a novel centrality metric for each protein in the network. This metric aims to capture both the individual interactions of the protein, as well as the interactions of its open neighborhood, when disregarding neighboring nodes that are connected to one another. We refer to this centrality as *star centrality*.

1.1 Outline

We first provide a literature review on protein-protein interaction networks and protein essentiality, along with the definition of "party" and "date" hubs, and centrality. In Section 2, we present the basic notation we will be using throughout the paper, define the problem, and provide its computational complexity. Then, Section 3 focuses on our mathematical programming framework; in the same section we propose greedy heuristic approaches to tackling the problem faster and provide their approximation guarantees. Section 4 presents our computational study on six protein-protein interaction networks, namely *Saccharomyces Cerevisiae* (yeast), *Helicobacter Pyloris*, *Staphylococcus Aureus*, *Salmonella Enterica CT18*, *C. Elegans*, and *Homo Sapiens* (human). The performance of the approximation algorithms is also contrasted to the exact solution. We conclude this work with our observations and our insights.

1.2 Protein-protein interaction networks

Protein-protein interaction networks (PPIN) have become, mostly over the last decade, an important point of discussion for many disciplines in their quests to better understand and analyze how and why proteins interact with one another. As proteins are fundamental entities that control numerous biological activities, information on how they bind and interact to perform said activities is an important scientific exercise that can bring to light insight into cell mechanisms.

The first step towards creating a PPIN is to discover pairs of proteins that interact with each other. This is typically performed experimentally using two-hybrid screening or yeast two-hybrid (Y2H) (Sardiu and Washburn

2011), or coaffinity purification and mass spectrometry (AP/MS) systems (Teng *et al.* 2014). After further analysis on the individual interactions, a collection of them comprises the overall network that can be used. PPINs are now readily available from many different databases, such as the ones by Xenarios *et al.* (2000), Zanzoni *et al.* (2002), Pagel *et al.* (2005), Franceschini *et al.* (2013), Chatr-Aryamontri *et al.* (2013), among others. Even though it has been observed that such networks are not without errors (Legrain and Selig 2000, Hart *et al.* 2006), it is still valuable to analyze them using complex network analysis as they provide us with interesting information on how proteins work.

The PPIN that has been most well-studied is the one belonging to *C. Elegans*. The reason behind it is mostly the fact that there exist a great deal of similarity between human and *C. Elegans*; more than 50% of the genes present in *C. Elegans* are also present in their homologue form in the human genome, as observed by Kamath *et al.* (2003). Another well-studied PPIN belongs to the *S. Cerevisiae* organism; interestingly, this dataset has also been employed to showcase the effectiveness of node criticality (Veremyev *et al.* 2015).

1.3 Essential proteins and hubs

A fundamental question in the analysis of PPINs (as well as in general biological networks) is whether there exist proteins (nodes) that significantly alter its functionality (or, even result in lethality). A protein is said to be *essential* or *lethal* when, if absent, it causes the biological cell to die (Kamath *et al.* 2003) or prevents it from reproducing properly. The study of essential proteins was and still is performed experimentally; however, those experiments tend to be expensive, both resource- and time-wise. An example of such a technique is conditional gene knockout, a technique in which a specific gene is removed from a tissue (Skarnes *et al.* 2011).

It has been observed that the study of protein essentiality can be targeted to only a select number of proteins (or, equivalently, proteins can be discarded from contention) using quantitative methods on the rapidly increasing, available data. The interested reader is referred to the excellent work by Zotenko *et al.* (2008) as well as the *centrality-lethality rule*, which was one of the first attempts to study essentiality as a function of network topology (Jeong *et al.* 2001). Nowadays, with the availability of vast amounts of proteomic data, information on essentiality of proteins is also increasing:

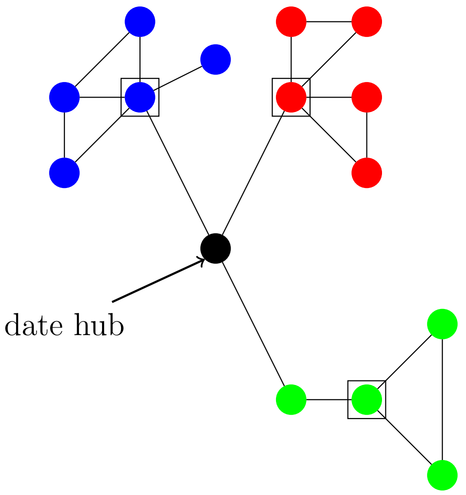


Figure 1: An example of how "party" and "date" hubs would appear as in a PPIN. The nodes of blue, red, and green color represent three different structures/complexes in a generated PPIN. The "party" hubs are marked with a square, while the "date" hub is annotated on the Figure. Observe that the "date" hub in the middle has a smaller number of interactions (smaller degree) than some of the "party" hubs in this example.

for instance, we refer the reader to the curated Database of Essential Genes, or DEG (Zhang *et al.* 2004, Zhang and Lin 2009, Luo *et al.* 2013).

A *hub* is defined as a protein with many interactions. Seeing as this definition is open-ended, some researchers use different threshold values for the number of interactions. For instance, in the fundamental contribution by Han *et al.* (2004), a hub is defined as a protein with more than 5 interactions. Therein, the authors also investigate another, very important protein characterization, one between proteins that interact with all their neighbors simultaneously and proteins that interact with their partners in different times and/or locations, also referred to informally as "party" and "date" hubs, respectively. More formally stated, "party" hubs show high co-expression with their partners, while "date" hubs the opposite. An example of how the definition of "party" and "date" hubs would look like in a toy network is shown in Figure 1.

This computational discovery has been met with scrutiny by the scientific community and has led to a general debate on whether this classification of proteins actually helps us decode the proteome (Mirzarezaee *et al.* 2010). In general, though, this hypothesis has led to significant interest in connecting graph theoretic notions to PPIN analysis (e.g., the works by Agarwal *et al.* (2010) and Gursoy *et al.* (2008), among others).

1.4 Centrality

Centrality is a fundamental concept in network analysis, typically tied to the topological "importance" of a network element. As such, it is also a well-studied topic, starting from the contributions of Bavelas (1948, 1950), Leavitt (1951), Sabidussi (1966), and Freeman (1979) earlier, and reaching out to more recent contributions (e.g., Borgatti and Everett (2006), Koschützki *et al.* (2005), Boldi and Vigna (2013) for an extensive review of centrality indices). This list is by no means exhaustive, as it is a topic that has attracted scientific interest from multiple and diverse groups of researchers.

Basic centrality measures that have been proposed over the years can be categorized based on their local or global considerations. For example, node *degree* centrality, which represents the number of nodes adjacent to a given node i, is a local metric as it only considers the neighborhood of the node at hand. On the other hand, node *betweenness* centrality, a metric that can be defined as the fraction of the shortest paths from any two nodes in the network that use a node i as an intermediary, is global.

Indicatively, we present some basic centrality measures that have appeared in the literature over the years.

- *Degree* centrality: given a node i, its degree centrality is the number of nodes adjacent to i;

- *Closeness* centrality: given a node i, its closeness centrality is the maximum/average path length to every other node;

- *Betweenness* centrality: for any node i, it captures the fraction of the shortest paths connecting two other nodes in the network and use i. It is typically further divided into probabilistic (considering all geodesic paths), optimistic (considering only the geodesic path passing through i), and pessimistic (considering only the geodesic paths that do not use i);

- *Eigenvector* centrality: the main idea here is that the centrality of any node i is higher if they are connected to other highly centralized nodes. PageRank (Page *et al.* 1999) can be viewed as a special case of eigenvector centrality.

A specific extension that is of interest to us has to do with *group* centrality. Recently, we have seen more work that focuses on extending centrality notions to a group of nodes in the network (Everett and Borgatti 1999, 2005, Borgatti 2006). This extension enables us with notions of endogenous and exogenous centrality (Everett and Borgatti 2010), where a network property is taken and measured after node/edge deletion, and also provides us with a tool to consider clusters of nodes and figure out their topological importance. An integer programming formulation for detecting informal, cohesive groups with high and low centrality was presented by Vogiatzis *et al.* (2015).

Centrality has been a recurring theme in the study of biological networks, and, more specifically, PPINs. It is generally assumed that "important" areas of a network will also prove to be more centralized. The goal of using centrality metrics in this context is to find what the relationship is between the topology and the functionality of PPINs. As mentioned before, an attempt to draw conclusions for protein essentiality and centrality dates back to the contribution by Jeong *et al.* (2001). More recently, Yu *et al.* (2007) investigated the importance of bottleneck proteins in PPINs. Their results seem to support the idea that bottlenecks are indeed very important in recognizing essential proteins as well as "date" hubs. Numerous studies have proposed relationships between nodal centrality metrics and essentiality in PPINs (Joy *et al.* 2005, Hahn and Kern 2005, He and Zhang 2006). In our work, we investigate these claims and contrast the performance of said centrality metrics to our proposed methodology.

While centrality has indeed proven an important characteristic of PPINs, there are some caveats with the approaches currently in practice. First, assigning importance to a single protein (resp. interaction), instead of a set of proteins (resp. interactions) tends to favor those proteins that participate in large complexes. Secondly, PPINs are still not error-free (Hart *et al.* 2006); assuming complete information can lead to significant misattributions of importance. Last, some proteins that present low co-expression with their interacting partners would be disregarded by such metrics even though they might have a significant role in coordinating different complexes (e.g., "date" hubs). We will show that our proposed approach alleviates all three of these issues. We can now proceed to formally state the notation and the definition of the problem in the next section.

2 Fundamentals

Let $G(V, E)$ represent a simple, undirected graph with a nodeset V of size $|V| = n$ nodes and an edgeset $E \subset V \times V$ of size $|E| = m$. We say that two nodes $i, j \in V$ are connected by an edge if the adjacency matrix entry $a_{ij} = 1$; otherwise we have that $a_{ij} = 0$. Seeing as the graphs considered here are undirected, the adjacency matrix is symmetric. We further consider a positive weight parameter on the edges of the graph, $w_e : E \mapsto \mathbb{R}, \forall e = (i, j) \in E$. Furthermore, the open neighborhood of a node $i \in V$ is defined as $N(i) = \{j \in V : (i, j) \in E\}$; similarly, the closed neighborhood of a node i is defined as $N[i] = N(i) \cup \{i\}$. The notion of (open) neighborhood is sometimes generalized to include nodes that are reachable within at most k hops. This neighborhood is represented here by $N^k(i)$: for example, the complete set of nodes reachable by $i \in V$ within at most 2 hops would be denoted as $N^2(i)$. Using the above definitions, node degree centrality can be easily represented as

$$\mathcal{C}^d(i) = |N(i)|.$$

We also define the subgraph induced by a set of nodes S, $G[S]$ as the subgraph of G with a vertex set $V[G[S]] = S$ and an edge set $E[G[S]] = \{(i, j) \in E : i, j \in S\}$. We further say that a set of nodes S forms an *induced star* if the induced subgraph of S has exactly one node of degree $|S| - 1$ and $|S| - 1$ nodes of degree 1.

2.1 Problem definition

In this work, we define a centrality measure that incorporates information from the centrality of the open neighborhood, instead of relying solely on the considered node. More specifically, we focus on degree centrality:

Definition 1. *The star degree centrality of a node i is the degree centrality of the induced star S centered at i that produces the maximum open neighborhood size of S.*

Formally, this can be expressed as in (1).

$$\mathcal{C}^s(i) = \max\{|N(S)| : S \subseteq V \textit{ forms an induced star centered at } i \in V\} \quad (1)$$

As an example, let us return to the graph of Figure 1. Consider first the portrayed date hub in the middle. The induced star, centered at the date hub, that produces the maximum open size neighborhood would be either the set S consisting of the date hub, and the blue and red party hubs (with a value of 9 nodes adjacent to S), or it could also include the green node in the lower right connection of the date hub (this set would also have a star centrality value of 9). In contrast, consider the blue party hub, which originally has the biggest degree (along with the red party hub). Its star centrality now can be shown to be found when considering the induced star formed by the blue party hub and the date hub itself (having a value of 6). Last, let us consider the green party hub. For that node, we can easily verify that its degree and star centrality match (and are equal to 3).

2.2 Complexity

In this subsection, we provide the computational complexity of the problem of detecting the node of maximum star degree centrality. We first give the decision version of the problem at hand.

Definition 2 (STAR DEGREE CENTRALITY). *Given a graph $G(V, E)$ and an integer k, does there exist an induced star S centered at any node $i \in V$ such that $|N(S)| \geq k$?*

We proceed to derive the complexity of the problem using the well-known $\mathcal{NP}$-complete problem, INDEPENDENT SET.

Definition 3 (INDEPENDENT SET). *Given a graph $G(V, E)$ and an integer k, does there exist a set $S \subseteq V$ such that $|S| \geq k$ and for any two nodes $i, j \in S$, $(i, j) \notin E$?*

Theorem 1. *STAR DEGREE CENTRALITY is $\mathcal{NP}$-complete.*

Proof. First of all, it is easy to verify that the problem is in NP. Given a set of nodes $S \subseteq V$, we can verify that S forms an induced star (one center with degree of $|S| - 1$ and no connections between leafs) and that $|N(S)| \geq k$ in polynomial time.

Now, consider an instance of INDEPENDENT SET $< G, k >$. We construct an instance of STAR DEGREE CENTRALITY $< \hat{G}, \ell >$ as follows. The graph $\hat{G}$ is defined as:

$$V[\hat{G}] = \hat{V} = V \cup \{s\} \cup \left\{ \cup_{i=1}^{n} \{ \cup_{j=1}^{n} s_j^{(i)} \} \right\} \cup \left\{ \cup_{i=1}^{n^2} s_i^{(s)} \right\}$$

$$E[\hat{G}] = \hat{E} = E \cup \{ \cup_{i=1}^{n} (s, i) \} \cup \{ \cup_{i=1}^{n} \{ \cup_{j=1}^{n} (i, s_j^{(i)}) \} \} \cup \left\{ \cup_{i=1}^{n^2} (s, s_i^{(s)}) \right\}$$

The above imply that graph $\hat{G}$ has the nodes from G with their connections, a newly added node s that is connected to every other node in V, $n = |V|$ nodes for every node $i \in V$ ($s_j^{(i)}$, for $j = 1, \ldots, n$) that are connected to i, and n^2 nodes ($s_j^{(s)}$, for $j = 1, \ldots, n^2$) that are only connected to s. In total the new graph has $2n^2 + n + 1$ nodes and $2n^2 + m + n$ edges. Furthermore, let $\ell = k \cdot n$. An example of the reduction from INDEPENDENT SET to STAR DEGREE CENTRALITY can be found in Figure 2.

Let S be an independent set of size k in G. Then, consider the set of nodes $\hat{S} = S \cup \{s\}$. It is easily verified that $\hat{S}$ forms an induced star by construction and because S is an independent set. Furthermore, it is straightforward to see that $|N(\hat{S})| \geq k \cdot n + n^2$.

Now, assume that there exists no independent set of size k in G. For a contradiction, we assume that there exists an induced star $\hat{S} \subseteq \hat{V}$ such that $|N(\hat{S})| \geq k \cdot n + n^2$. First of all, we note that $s \in \hat{S}$: if not, then there can be no star using nodes from $\hat{V} \setminus \{s\}$ with an open neighborhood of size at least equal to n^2. Further, there exist at least k nodes from V in the star: once more, if that is not the case, then $n^2 \leq |N(\hat{S})| < n^2 + k \cdot n$. Last, observe that $\hat{S}$ is centered in s: assume for a contradiction that the star is instead centered at a node $i \in V$. Then, one of the two following cases has to hold:

(a) $\hat{S}$ contains s. This implies that no other node $j \in V$ can be in the star, as $(s, j) \in \hat{E}$;

(b) $\hat{S}$ contains at least k nodes in V. This implies that s cannot belong in $\hat{S}$, for the same reason as above.

In both cases, we observe that we reach a contradiction, hence $\hat{S}$ has to be an induced star centered at s. By construction, and since $\hat{S}$ forms an induced star with k nodes in V, there exists no edge connecting any two nodes in $\hat{S} \setminus \{s\}$, which shows that it actually forms an independent set of size k in G. This contradiction finishes the proof. □

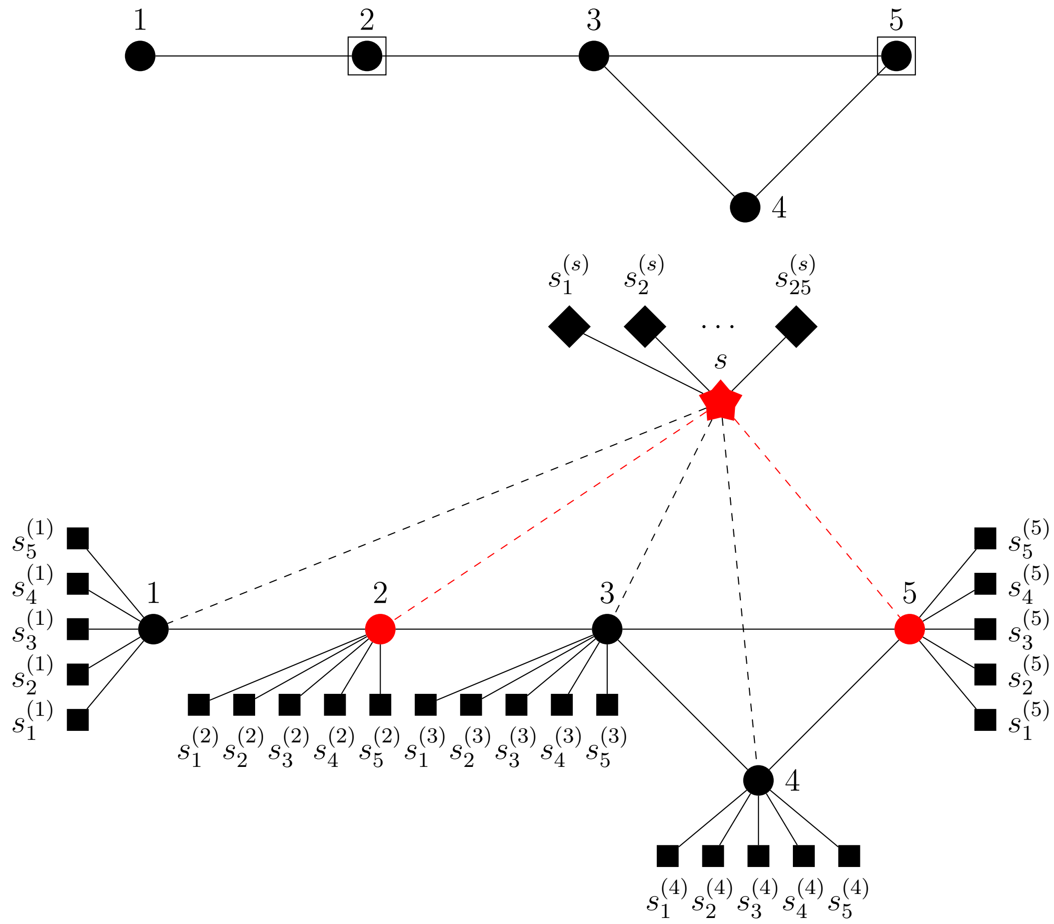

1
2
3
5
4
$s_1^{(s)}$
$s_2^{(s)}$
. . .
$s_{25}^{(s)}$
s
$s_5^{(1)}$
$s_4^{(1)}$
$s_3^{(1)}$
$s_2^{(1)}$
$s_1^{(1)}$
1
2
3
5
$s_5^{(5)}$
$s_4^{(5)}$
$s_3^{(5)}$
$s_2^{(5)}$
$s_1^{(5)}$
$s_1^{(2)}$ $s_2^{(2)}$ $s_3^{(2)}$ $s_4^{(2)}$ $s_5^{(2)}$ $s_1^{(3)}$ $s_2^{(3)}$ $s_3^{(3)}$ $s_4^{(3)}$ $s_5^{(3)}$
4
$s_1^{(4)}$ $s_2^{(4)}$ $s_3^{(4)}$ $s_4^{(4)}$ $s_5^{(4)}$

Figure 2: An example of the reduction for a graph G. Above, it is easy to see that the selected nodes form an INDEPENDENT SET of size 2. Below, there exists an induced star S such that $|N(S)| \geq n^2 + k \cdot n = 25 + 10 = 35$. The star represents the new added node, the squares are the n nodes in $\hat{G}$ connecting to every node in V, while the diamonds the n^2 nodes connected to the newly added star node.

2.3 Extensions

2.4 Extensions

It can also be shown that the star centrality function is submodular.

Theorem 2. *The function $f(S) = \{|N(S)| : S \text{ forms an induced star}\}$ is submodular.*

Proof. Let S_1, S_2 be two induced stars such that $S_1 \subseteq S_2$. Also, consider a node $u \in V \setminus S_2$. Then, we have that:

$$f(S_1 \cup \{u\}) - f(S_1) = -1 + |N(u)| - |N(u) \cap N[S_1]|$$
$$f(S_2 \cup \{u\}) - f(S_2) = -1 + |N(u)| - |N(u) \cap N[S_2]|$$

It is clear that $|N(u) \cap N[S_1] \leq |N(u) \cap N[S_2]|$, as $N[S_1] \subseteq N[S_2]$, and hence, $f(S_1 \cup \{u\}) - f(S_1) \geq f(S_2 \cup \{u\}) - f(S_2)$. □

Unfortunately though, the star centrality function is not monotone; consider a node with no neighbors other than to a designated “center”. Then, that node can be used as a leaf to a star, however it would increase its open neighborhood size by 1. An indicative counterexample is presented in Figure 3. This implies that we cannot easily use a simple greedy approach to approximate the optimal solution. We do though provide a different greedy mechanism study in a subsequent section.

3 Mathematical Formulation and Approximation Algorithms

In this section, we provide a mathematical formulation for our problem, followed by two approximation algorithms. First, let us define the following decision variables:

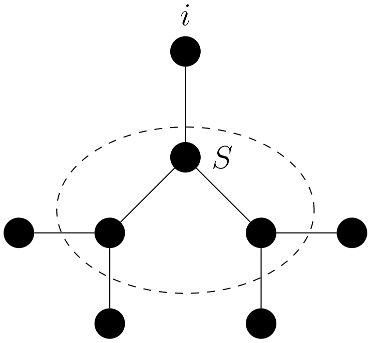


Figure 3: A counterexample of the monotonicity of the star centrality function. As can be easily seen, the open neighborhood size of the star S is decreased by 1 when considering the star $S \cup \{i\}$.

$$x_i = \begin{cases} 1, & \text{if node } i \in V \text{ is the center of the star} \\ 0, & \text{otherwise.} \end{cases}$$

$$y_i = \begin{cases} 1, & \text{if node } i \in V \quad \textit{is in the star} \\ 0, & \text{otherwise.} \end{cases}$$

$$z_i = \begin{cases} 1, & \text{if node } i \in V \text{ is adjacent to a node in the star} \\ 0, & \text{otherwise.} \end{cases}$$

3.1 Mathematical Formulation

The integer programming formulation for detecting the induced star of maximum degree centrality is presented in (2)–(7).

$$
\begin{aligned}
\text{IP:}\ \max \sum_{i\in V} z_i & & & \quad (2)\\
s.t.\ y_i + z_i \le 1, & & \forall i \in V & \quad (3)\\
z_i \le \sum_{j\in N(i)} y_j, & & \forall i \in V & \quad (4)\\
y_i \le \sum_{j\in N[i]} x_j, & & \forall i \in V & \quad (5)\\
y_i + y_j \le 1 + x_i + x_j, & & \forall (i,j) \in E & \quad (6)\\
\sum_{i\in V} x_i = 1, & & & \quad (7)\\
x_i, y_i, z_i \in \{0,1\}, & & \forall i \in V. & \quad (8)
\end{aligned}
$$

Clearly, our objective is to maximize the size of the open neighborhood of the star, as shown in (2). Then, (3) ensures that no node is allowed to be both in the star and in its open neighborhood. Constraint families (4) and (5) are similar in nature and enforce which nodes are adjacent to the star, and which nodes are adjacent to the center and, as such, can be considered for addition to the star. Moreover, no two leafs are allowed to be connected, as per constraint (6). Last, we are only looking for one star, enforced with (7), and all of our decision variables are binary.

We can also consider the problem of detecting the star centrality of a given node $u \in V$, as shown in (9)–(14).

$$
\begin{aligned}
\text{IP}(u)\text{:}\ \max \sum_{i\in V} z_i & & & \quad (9)\\
s.t.\ y_i + z_i \le 1, & & \forall i \in V & \quad (10)\\
y_i \le a_{iu}, & & \forall i \in V \setminus \{u\} & \quad (11)\\
z_i \le \sum_{j:(i,j)\in E} y_i, & & \forall i \in V & \quad (12)\\
y_i + y_j \le 1, & & \forall (i,j) \in E : i \ne u, j \ne u & \quad (13)\\
y_i, z_i \in \{0,1\}, & & \forall i \in V. & \quad (14)
\end{aligned}
$$

The objective function (shown at (9)), as well as constraint families (10), (12) and the variable restrictions at (14) are identical to the previous model.

However, note that we no longer need to consider a decision variable for the center of the star, as it is known to be node $u \in V$. Hence, we can add constraints (11) that only consider the nodes that are adjacent to u as candidates to be in the star, and modify constraint (13) to only consider the connections that do not include the star center. As a reminder, a_{ij} is the adjacency matrix entry that represents the connection between nodes i and j.

3.2 Greedy algorithms

As shown earlier, we cannot unfortunately claim monotonicity for the star centrality function. Hence, deriving an approximation ratio from simply applying a greedy algorithm scheme is not straightforward. However, we can still show that the greedy algorithm, presented in Algorithm 1 has an approximation guarantee of $O(\Delta)$, where Δ is the maximum degree in the network. First, let us introduce for simplicity a function $f_1(S,k)$ to capture the "gain" of adding a node k to a star S, assuming of course that $S \cup \{k\}$ remains an induced star. We note that for this function we have that $f_1(S,k) \geq -1$.

$$f_1(S,k) = |N(S \cup \{k\})| - |N(S)|.$$

Theorem 3. *Let $i \in V$, with a degree of δ, be the node whose star centrality we are interested in finding. Then, the simple greedy algorithm has an an approximation ratio of $O(\delta)$.*

Proof. At each iteration of the while loop, the greedy algorithm looks at the candidate nodes (set $\{j \in N(i) \setminus S : (k,j) \notin E, \forall k \in S\}$), and selects to add the one that is adjacent to the maximum number of not already covered nodes. In the worst case, the greedy algorithm terminates after the first iteration, and that only happens when the greedily selected node $u \in N(i)$ which adds $\alpha = |N(u) \setminus N[i]|$ is connected to every other node in $N(i)$. Let OPT be the optimal value and z_{greedy} the value obtained by applying the simple greedy approach. Then, we have that

$$OPT \leq (\delta - 1) \cdot (\alpha - 1) + 1 \leq (\delta - 1) \cdot \alpha \tag{15}$$

$$z_{greedy} \geq \alpha + \delta - 1 \geq \alpha \tag{16}$$

Algorithm 1: Simple Greedy.

```
function SimpleGreedy ($i$);
Input  : A node $i \in V$
Output: An induced star $S$ centered at $i$
$candidates \leftarrow N(i)$;
$S \leftarrow \{i\}$;
while $candidates \neq \emptyset$ do
    for $k \in candidates$ do
        if $f_1(S, k) <= 0$ then
            $candidates \leftarrow candidates \setminus \{k\}$;
        end
    end
    if $candidates \neq \emptyset$ then
        $j \leftarrow \arg\max_k \{f_1(S, k) : k \in candidates\}$;
        $S \leftarrow S \cup \{j\}$;
        $candidates \leftarrow candidates \setminus \{N[j]\}$
    end
end
return $S$
```

From (15) and (16), we obtain the approximation guarantee as

$$\frac{OPT}{z_{greedy}} \leq \frac{(\delta - 1) \cdot \alpha}{\alpha} = \delta - 1 = O(\delta). \tag{17}$$

□

Figure 4 shows an example of the worst-case performance. Let us now propose a different greedy-based heuristic algorithm and show its approximation ratio. Let S^i be again an induced star centered at i and define function $f_2(S^i, k)$ as:

$$f_2(S^i, k) = \sum_{j:(i,j) \in E, (k,j) \in E} | \big(N(S^i \cup \{j\}\big) \setminus N(S^i)|.$$

This function captures the potential increase in the size of the open neighborhood that we would be losing since nodes j and k cannot belong to the

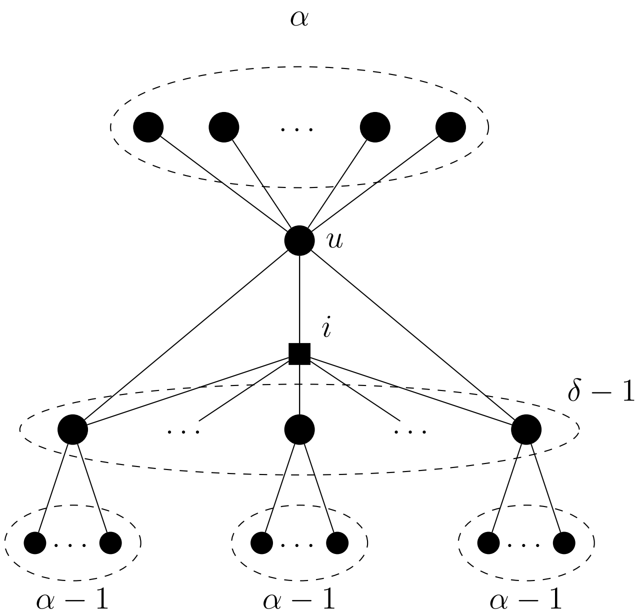


Figure 4: An example of the worst-case behavior guarantee of the Simple Greedy approach. In this case adding u to the star centered at i results in a star centrality of $\alpha + \delta - 1$, while adding every other neighbor of i to the star would result in $(\delta - 1) \cdot (\alpha - 1) + 1$.

star simultaneously. Note here that $f_2(S^i, k) = 0$ implies either that node k is connected to no other potential leaf of the star, or that all other candidates connected to k add no uncovered nodes to the star. Now, consider the greedy approach shown in Algorithm 2. We show its approximation ratio in Theorem 4; to do that, we first provide two lemmata.

Algorithm 2: Ratio-based Greedy.

```
function RatioGreedy ($i$);
Input  : A node $i \in V$
Output: An induced star $S$ centered at $i$
$S^i \leftarrow \{i\}$;
$candidates_1 \leftarrow \{k \in N(i) : f_2(S^i, k) = 0\}$;
$candidates_2 \leftarrow N(i) \setminus candidates_1$;
while $candidates_1 \neq \emptyset$ or $candidates_2 \neq \emptyset$ do
    if $candidates_1 \neq \emptyset$ then
        $j \leftarrow \arg\max_k \{f_1(S^i, k) : k \in candidates_1, f_1(S^i, k) > 0\}$;
        $S^i \leftarrow S^i \cup \{j\}$;
        $candidates_1 \leftarrow candidates_1 \setminus \{j\}$;
    else
        $j \leftarrow \arg\max_k \{\frac{f_1(S^i,k)}{f_2(S^i,k)} : k \in candidates_2, f_1(S^i, k) > 0\}$;
        $S^i \leftarrow S^i \cup \{j\}$;
        $candidates_2 \leftarrow candidates_2 \setminus N[j]$;
    end
    for $k \in candidates_2$ do
        if $f_2(S^i, k) = 0$ then
            $candidates_2 \leftarrow candidates_2 \setminus \{k\}$;
            $candidates_1 \leftarrow candidates_1 \cup \{k\}$;
        end
    end
end
return $S$
```

Lemma 1. *Let $i \in V$, with a degree of δ, be the node whose star centrality we are interested in finding. Further, assume that for all nodes $k \in N(i)$, we have that $f_2(S^i, k) = 0$, that is there exists no connection between any two* *of them. Then, greedily selecting the node with maximum $f_1(S^i, k)$ has an approximation ratio of $O(\ln \delta)$.*

Proof. It can be seen that the above setup results in greedily solving a set cover problem with δ sets. The universe of elements to be covered is all nodes reachable within 1 or 2 hops from i, $N^2(i)$. Each set consists of the neighbors of i and their neighbors which belong to $N^2(i)$, that is $C_j = \{j, N(j) \cap N^2(i)\}, \forall j \in N(i)$. Since applying the greedy algorithm results in an $O(\ln n)$ approximation for the set cover and we have at most δ candidate nodes/sets, all of which can be selected at any point, as there exist no connections between them, the greedy algorithm would result in an $O(\ln \delta)$ approximation ratio, as far as the number of nodes added to the star is concerned. Let OPT_{SC} represent the optimal solution to the set cover problem above and z_{SC} the solution using the greedy algorithm. We then have that:

$$OPT = |N^2(i)| - OPT_{SC} \tag{18}$$

$$z_{greedy} = |N^2(i)| - z_{SC} \geq |N^2(i)| - \ln \delta \cdot OPT_{SC} \tag{19}$$

Combining (18) and (19), we obtain that:

$$\frac{OPT}{z_{greedy}} \geq \frac{|N^2(i)| - OPT_{SC}}{|N^2(i)| - \ln \delta \cdot OPT_{SC}} \geq \frac{1}{\ln \delta} \implies z_{greedy} \leq \ln \delta \cdot OPT. \tag{20}$$

The last inequality proves the Lemma.

□

Lemma 2. *Let $i \in V$, with a degree of δ, be the node whose star centrality we are interested in finding. Further, assume that for all nodes $k \in N(i)$, we have that $f_2(S^i, k) > 0$, that is each node is connected to at least one other in $N(i)$. Then, greedily selecting the node with maximum $\frac{f_1(S^i,k)}{f_2(S^i,k)}$ has an approximation ratio of $O(\sqrt{\delta})$.*

Proof. Similarly to the case in Theorem 3, the worst case behavior is observed when the algorithm terminates after adding only one node in the star. This can happen when the selected node is indeed adjacent to all other nodes in $N(i)$. Let β_j be the nodes adjacent to j that are not already in S or covered by S. Furthermore, let node u be connected to all other candidate nodes. We then have that:

$$a_u = \frac{f_1(S,u)}{f_2(S,u)} = \frac{\beta_u}{\sum\limits_{k \in N(i), k \neq u} \beta_k},$$

while, for the remaining nodes, $j \neq u$, we would have that:

$$a_j \leq \frac{\beta_j}{\beta_u}.$$

In the worst case, the remaining nodes can all be part of the same star (i.e., there exist no connections between them). Hence, to select node u using the ratio-based greedy approach we must have $a_u \geq a_j$, for all j, and assuming v is the nodes with maximum ratio when excluding u, we have that $a_u \geq a_v$. This implies:

$$\begin{aligned} a_u \geq a_v &\implies \frac{\beta_u}{\sum\limits_{k \in N(i), k \neq u} \beta_k} \geq \frac{\beta_v}{\beta_u} \implies \frac{\beta_u}{(\delta - 1) \cdot \beta_v} \geq \frac{\beta_v}{\beta_u} \\ &\implies \beta_u^2 \geq (\delta - 1) \cdot \beta_v^2 \implies \beta_v \leq \frac{\beta_u}{\sqrt{\delta - 1}}. \end{aligned} \tag{21}$$

Hence, in the worst case, the greedy algorithm results in a solution of $\beta_u + \delta - 1$, while the optimal solution can be as big as $(\delta - 1) \cdot \frac{\beta_u}{\sqrt{\delta - 1}} + 1$. We finally get:

$$\frac{OPT}{z_{greedy}} \leq \frac{(\delta - 1) \cdot \frac{\beta_u}{\sqrt{\delta - 1}} + 1}{\beta_u + \delta - 1} \leq \frac{\sqrt{\delta - 1} \cdot \beta_u}{\beta_u} = O(\sqrt{\delta}). \tag{22}$$

□

Theorem 4. *Let $i \in V$, with a degree of δ, be the node whose star centrality we are interested in finding. Then, the ratio-based greedy algorithm has an approximation ratio of $O(\sqrt{\delta})$.*

Proof. The algorithm is divided into two phases: in the first phase, the node with the maximum ratio is selected, while in the latter one, we choose the node with the maximum number of uncovered neighbors.

Let OPT_1 and OPT_2 represent the optimal solutions obtained from each phase. Then, $OPT \leq OPT_1 + OPT_2$. Similarly, let z_1 and z_2 be the solutions obtained from each phase of the greedy algorithm; it is easy to see that $z_{greedy} = z_1 + z_2$. From the previous lemmata, we have that:

$$OPT_1 \leq O(\sqrt{\delta}) \cdot z_1 \tag{23}$$

$$OPT_2 \leq O(\ln \delta) \cdot z_2. \tag{24}$$

Combining, we get that

$$\begin{aligned} \frac{OPT}{z_{greedy}} \leq \frac{OPT_1 + OPT_2}{z_{greedy}} \leq \frac{O(\sqrt{\delta}) \cdot z_1 + O(\ln \delta) \cdot z_2}{z_1 + z_2} \leq \\ \leq \frac{O(\sqrt{\delta}) \cdot (z_1 + z_2)}{z_1 + z_2} = O(\sqrt{\delta}). \end{aligned} \tag{25}$$

□

4 Computational results

In this section, we present our experimental setup, the data used, and analyze and interpret the results obtained. Our goal is to portray how star centrality behaves and performs when put to the test against other popular centrality metrics in PPIN analysis.

4.1 Experimental setup

All numerical experiments were performed on a quad-core Intel i7 at 2.8 GHz with 16 GB of RAM. The codes were written in Python and C++ and, where needed, the Gurobi 6.50 solver (Gurobi Optimization 2015) was used to solve the optimization problems. Data on protein interactions for different organisms was obtained by STRING v. 10.0 (Szklarczyk *et al.* 2014). More specifically, we used the datasets of *Saccharomyces Cerevisiae*, *Helicobacter Pyloris*, *Staphylococcus Aureus*, *Salmonella Enterica CT18*, *Caenorhabditis Elegans*, and *Homo Sapiens*. Essentiality for proteins was found using the databases for the above organisms as curated in DEG 10 (Luo *et al.* 2013).

The PPINs for all organisms were created as follows. For each protein in the database, a node was created and was connected to all other proteins-nodes that they shared an interaction. Then, all interactions-edges with an interaction score that was below a threshold were removed. Seeing as the maximum interaction score was 1000, the threshold score selected for presentation in this study was 600 (60% interaction score). In this fashion, we were able to create a network where all known centrality metrics can be captured given the computational power. The network was further broken down into its connected components with each component being independently analyzed, without loss of generality.

Last, nodal metrics of centrality (*degree*, *closeness*, *betweenness*, *eigenvector*) were computed with a Python implementation, using NetworkX 1.9 (Hagberg *et al.* 2008). On the other hand, *star centrality* calculations were performed on the same networks with a C++ implementation. For smaller scale networks an exact solution was found for every node; however, for large-scale PPINs, such as the *Homo Sapiens* proteomic data, obtaining an exact solution proved too difficult a feat given the memory restrictions. In such cases, an approximate solution was obtained using greedy Algorithm 2.

4.2 Analysis

After obtaining all the metrics for the PPIN in consideration, we calculated the ratio of essential proteins found in the top k and the bottom k proteins (as ranked by each centrality metric). The bounds for each analysis are shown in Table 1 and were based on the total number of essential proteins present in the PPIN for each organism. As an example, for *Helicobacter Pyloris*, the number of essential proteins found in the PPIN was 435, and hence the top 500 proteins were investigated. Observe that the closer a metric gets to 100%, the more accurately it detects essential proteins.

Organism	$\|V\|$	$\|E\|$	Essential Proteins	Top k	Bottom k
Saccharomyces Cerevisiae	5414	84407	1221	1000	500
Helicobacter Pyloris	1521	17792	431	500	500
Staphylococcus Aureus	2521	16857	314	400	400
Salmonella Enterica CT18	4274	40165	543	500	500
Caenorhabditis Elegans	7608	93028	492	500	500
Homo Sapiens	12466	236631	1435	1500	1000

Table 1: Details of the PPIN and the bounds selected for each organism analysis.

For each organism then, we provide two Figures: one representing the

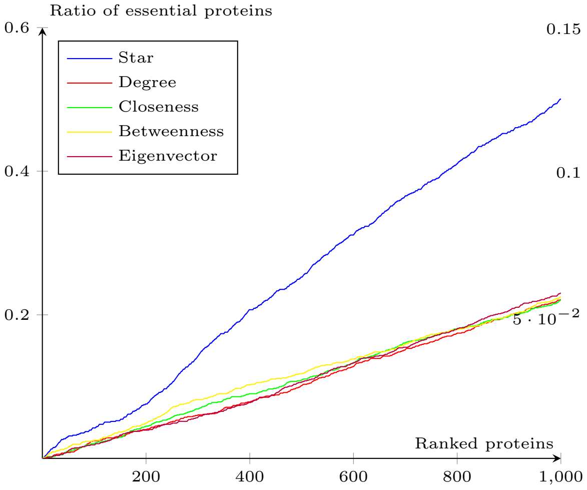


Figure 5: The ratio of essential proteins detected in the ranked *top* k proteins according to each metric for the *Saccharomyces cerevisiae* organism (yeast).

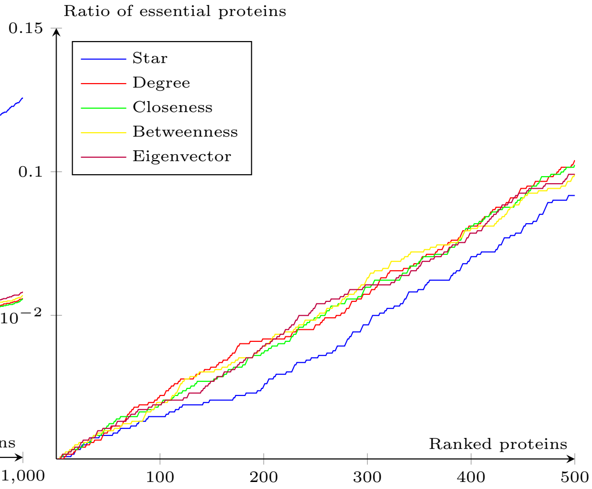


Figure 6: The ratio of essential proteins detected in the ranked *bottom* k proteins according to each metric for the *Saccharomyces cerevisiae* organism (yeast).

performance over the top k proteins, and a second one over the bottom k proteins. Note that for the first representation, the higher the ratio is then the better that metric is said to perform. The opposite is true for the second representation as a metric is said to perform better if the ratio is smaller. For example, consider Figures 5 and 6 that show our results for *Saccharomyces Cerevisiae*: the star centrality metric is impressively outperforming every other considered nodal centrality metric with a final performance of having 50.1% of all essential proteins within the top 1000. Note that the maximum that could be achieved here would be 81.9%, making the effective detection rate equal to 61.17%. On the contrary, the other centrality metrics are almost indistinguishable and achieve a final performance of 22.11%, 22.03%, 22.52%, and 23.01% for degree, closeness, betweenness, and eigenvector centrality, respectively. In the bottom 500 proteins, star centrality is still performing better, albeit less impressively, achieving a final score of 9.17%, as compared to the final scores of 10.4%, 10.24%, 9.91%, and 9.91%.

In the case of the *Helicobacter Pyloris* organism, shown in Figures 7 and 8, the situation is similar. Star centrality achieves a final score of detecting

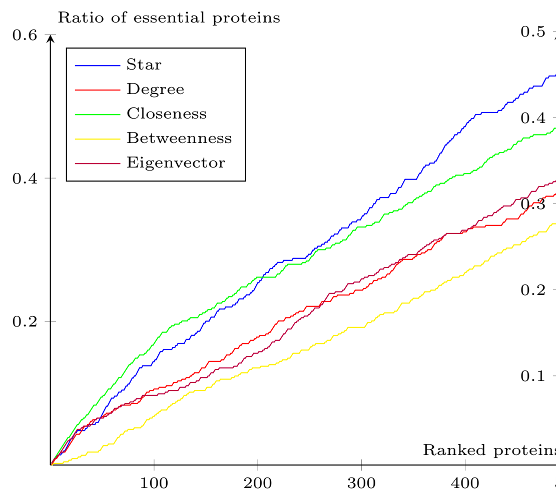


Figure 7: The ratio of essential proteins detected in the ranked *top k* proteins according to each metric for the *Helicobacter Pyloris* organism.

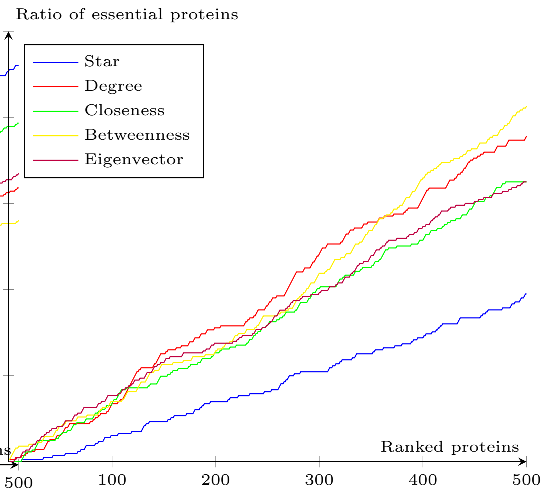


Figure 8: The ratio of essential proteins detected in the ranked *bottom k* proteins according to each metric for the *Helicobacter Pyloris* organism.

55.65% within the top 500 proteins, as opposed to 38.6% for degree centrality, 47.63% for closeness centrality, 34.09% for betweenness centrality, and 40.63% for eigenvector centrality. Considering the performance over the least well ranked proteins, it is easier to see that star centrality is best at not ranking highly non-essential proteins, achieving a final score of 19.49%, while the scores for the other centrality metrics are significantly higher at 37.82%, 32.51%, 41.31%, and 32.51%.

Continuing with the *Staphylococcus Aureus* organism (Figures 9 and 10), the same pattern is again seen. Star centrality consistently outperforms the other nodal metrics, and its accuracy is much higher at any given step in the analysis. Overall, the final star centrality score is 65.61%, which easily outperforms the final scores of the other centrality metrics, 40.21%, 38.14%, 39.18%, and 34.02%, respectively. Similarly, when considering the bottom 400 proteins, we obtain a final score of 7.96% for star centrality, as compared to the very high 38.14%, 45.36%, 37.11%, and 29.90% for the remaining centrality metrics.

Next, we continue our analysis with the *Salmonella Enterica subspecies CT 18* organism where the star centrality metric performs almost twice as

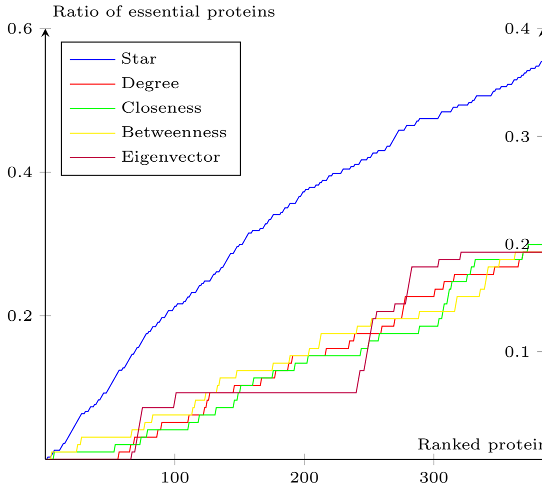


Figure 9: The ratio of essential proteins detected in the ranked *top* *k* proteins according to each metric for the *Staphylococcus aureus* organism.

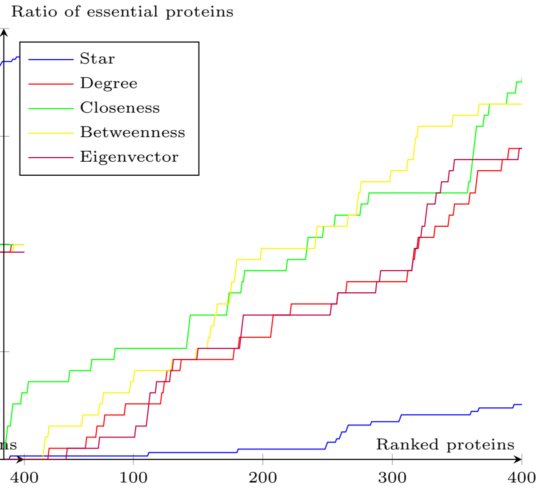
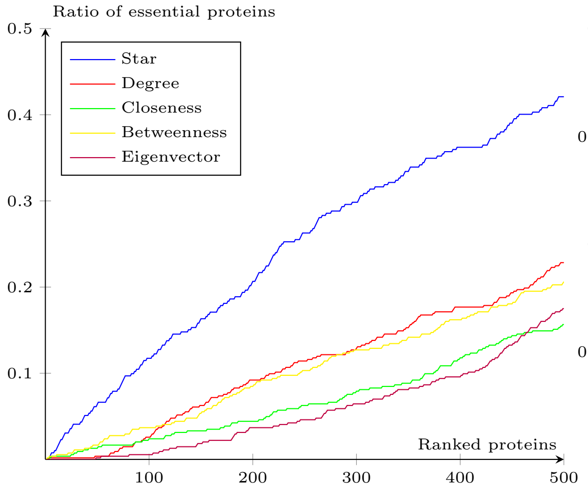


Figure 10: The ratio of essential proteins detected in the ranked *bottom* *k* proteins according to each metric for the *Staphylococcus aureus* organism.

well than any other centrality metric, with a final score of 42.09%, as can be seen in Figure 11. As a comparison, the score that is closest is the one of degree centrality (22.84%), while closeness, betweenness, and eigenvector centrality are at 15.65%, 20.63%, and 17.5%, respectively. When considering the bottom 500 proteins in Figure 12, once more star centrality with a score of 9.18% misclassifies less essential proteins than the other centrality metrics at 17.68%, 16.43%, 18.78%, and 19.71%.

As mentioned in the introduction, the *C. Elegans* organism is of particular interest as it shares common or homologue proteome to humans. Interestingly, for both organisms, star centrality and closeness centrality perform similarly. First, let us focus on Figures 13 and 14. As can be seen, star centrality barely outperforms closeness centrality (behaving similarly) with a final score of 47.29% compared to 41.58%. The other three metrics are far behind with scores of 32.19%, 30.98%, and 19.10% for degree, betweenness, and eigenvector centrality. As far as the bottom 500 ranked proteins are concerned, the corresponding scores are low and we note that a similarly low score is observed for the human proteome too. The scores are 3.38%, 5.40%, 4.35%, 4.89%, and 5.28% for the centrality metrics in the order presented in



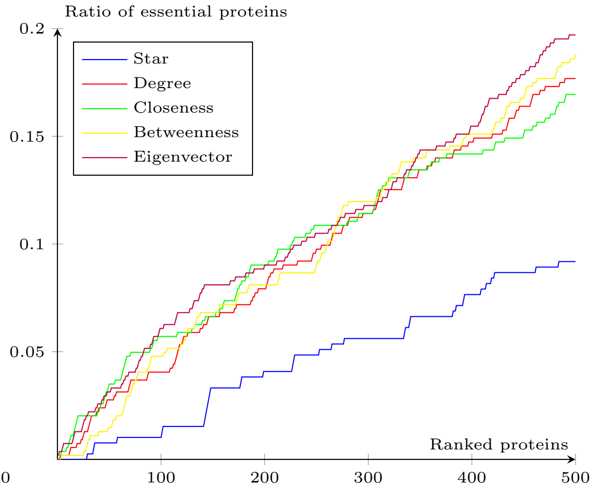


Figure 11: The ratio of essential proteins detected in the ranked *top* k proteins according to each metric for the *Salmonella enterica CT18* organism.

Figure 12: The ratio of essential proteins detected in the ranked *bottom* k proteins according to each metric for the *Salmonella enterica CT18* organism.

the Figure legends.

Last, we turn our attention to the biggest dataset available in this study. The *Homo Sapiens* proteome is very large-scale, and as such, it is more difficult to exactly calculate the star centrality metric for each available protein. Instead, we employ the ratio-based greedy algorithm (shown in Algorithm 2). Despite this fact, the star centrality ranking continues to perform best in both the top (Figure 15) and the bottom k proteins (Figure 16). The final scores are at 35.96% for star centrality, 21.05% for degree centrality, 30.24% for closeness centrality, 27.74% for betweenness centrality, and 21.11% for eigenvector centrality. When considering the bottom 1000 proteins, star centrality performs as well, misclassifying only 2.09% of essential proteins, as compared to the similarly low 3.9%, 2.79%, 3.83%, and 3% for the other centrality metrics.

4.3 Greedy Algorithm Analysis

In this subsection, we compare the performance of the two approximation algorithms in practice, using the same PPINs. The results are summarized

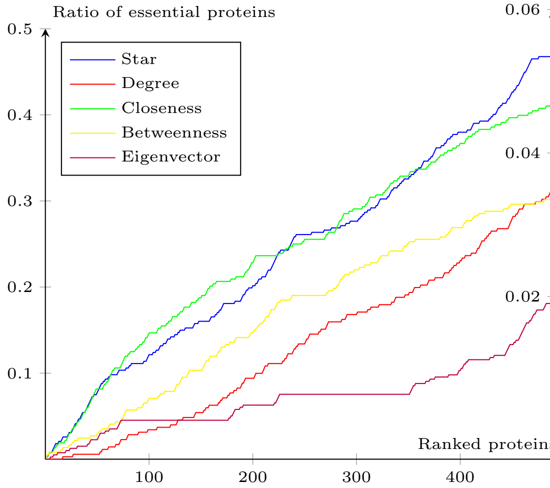


Figure 13: The ratio of essential proteins detected in the ranked *top* k proteins according to each metric for the *C. Elegans* organism.

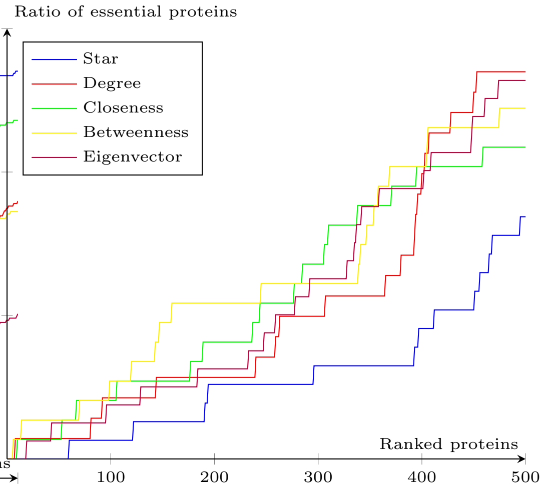


Figure 14: The ratio of essential proteins detected in the ranked *bottom* k proteins according to each metric for the *C. Elegans* organism.

in Tables 2 and 3. Note that for *Homo Sapiens*, the exact optimal solution was not found (see previous subsection), and hence the approximation ratios are not reported. We make the following observations. First, Algorithm 2 provides a better solution for every protein in every PPIN when compared to Algorithm 1. On average though, as can be seen in Table 2, both algorithms perform similarly well, finding the optimal solution in the majority of proteins.

	Average Approximation		Minimum Approximation		Optimal Found	
Organism	Simple	Ratio-based	Simple	Ratio-based	Simple	Ratio-based
Saccharomyces Cerevisiae	0.979	0.985	0.024	0.560	0.757	0.767
Helicobacter Pyloris	0.964	0.967	0.371	0.543	0.588	0.609
Staphylococcus Aureus	0.963	0.980	0.290	0.565	0.660	0.724
Salmonella Enterica CT18	0.966	0.969	0.241	0.559	0.628	0.644
Caenorhabditis Elegans	0.978	0.987	0.051	0.578	0.782	0.790

Table 2: Approximation ratio analysis for both Algorithms 1 and 2 for different PPINs. The last columns show the ratio of optimal solutions found.

More specifically, we note that in all organisms, *Ratio-based Greedy* always found a solution that was at least half as good as the optimal. On the

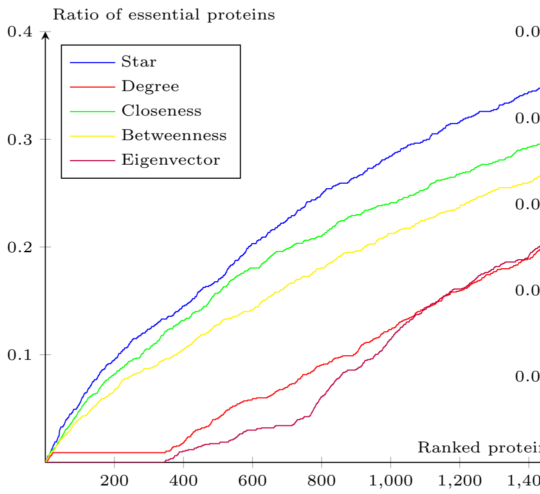


Figure 15: The ratio of essential proteins detected in the ranked *top* k proteins according to each metric for the *Homo Sapiens* organism (human proteome).

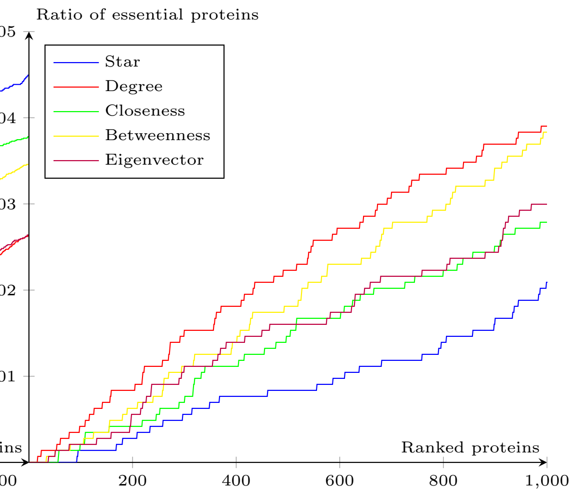


Figure 16: The ratio of essential proteins detected in the ranked *bottom* k proteins according to each metric for the *Homo Sapiens* organism (human proteome).

other hand, we note that there are occasions where the *Simple Greedy* fails to get a high quality solution and behaves close to its approximation guarantee. However, we can also observe that both approximation algorithms are able to find solutions that are very close to the optimal. In all organisms the solution obtained by either algorithm was on average as good as 96.3% of the optimal solution. This means that, even though in some cases the exact optimal is not found, the optimality gap is very small.

	Average Time			Maximum Time		
Organism	Simple	Ratio-based	Solver	Simple	Ratio-based	Solver
Saccharomyces Cerevisiae	0.074	0.122	0.151	18.841	174.160	438.155
Helicobacter Pyloris	0.025	0.036	0.049	0.498	2.476	4.176
Staphylococcus Aureus	0.026	0.038	0.047	1.303	3.188	3.924
Salmonella Enterica CT18	0.073	0.259	0.762	4.626	16.095	48.496
Caenorhabditis Elegans	0.131	0.853	1.781	21.147	194.716	1960.218

Table 3: Average and maximum computational times (in seconds) observed for the approximation algorithms and the Gurobi solver for different PPINs.

As far as our time study, shown in Table 3, is concerned, the main result is that, as expected, *Simple Greedy* outperforms both the more refined *Ratio-*

based Greedy and the Gurobi solver. This performance extends to both the average and the worst-case behavior of the three approaches.

5 Conclusions

In this work, we propose a new centrality metric, called *star centrality*, which aims to consider the connections of the "best" induced star centered at a node i. The problem was shown to be $\mathcal{NP}$-hard, however two approximation algorithms that perform efficiently, both as far as execution time and solution quality are concerned, were devised and implemented. The metric was then compared to traditional nodal centrality metrics in real-life protein-protein interaction networks, outperforming them in all instances; often significantly.

The implications from our work are two-fold. From a biological aspect, this metric provides researchers with a new and improved scoring scheme for ranking proteins and their interactions based on not only the proteins themselves, but also after considering their interacting partners. While our study is focusing on a specific type of clusters (induced stars), understanding how the new score works can prove valuable for developing other, group-based scoring/ranking schemes. Another important aspect of our contribution is that we were able to show that by considering groups of proteins we mitigate known problems with current large-scale proteome databases, improving the quality and robustness of the obtained scores.

We finally observe that the proposed metric does indeed take care of the three caveats mentioned earlier. First, this extension does not favor proteins that participate in a large number of interactions; instead it merely favors proteins that are located in "strategic", as far as the network topology is concerned, locations in the proteome. Secondly, if an error exists and an interaction is missing (or present, when it should not be), the effect it has in the metric is alleviated as a set of proteins is considered, instead of singleton proteins. Lastly, proteins with low co-expression that however serve to connect otherwise disconnected protein complexes will have a higher star centrality metric, helping in their identification, contrary to other centrality metrics in use for PPINs.

References

Sumeet Agarwal, Charlotte M Deane, Mason A Porter, and Nick S Jones. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol*, 6(6):e1000817, 2010.

Alex Bavelas. A mathematical model for group structures. *Human organization*, 7(3):16–30, 1948.

Alex Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *arXiv preprint arXiv:1308.2140*, 2013.

Stephen P Borgatti and Martin G Everett. A graph-theoretic perspective on centrality. *Social networks*, 28(4):466–484, 2006.

Stephen P Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, 2006.

Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, et al. The biogrid interaction database: 2013 update. *Nucleic acids research*, 41(D1):D816–D823, 2013.

Martin G Everett and Stephen P Borgatti. The centrality of groups and classes. *The Journal of mathematical sociology*, 23(3):181–201, 1999.

Martin G Everett and Stephen P Borgatti. Extending centrality. *Models and methods in social network analysis*, 35(1):57–76, 2005.

Martin G Everett and Stephen P Borgatti. Induced, endogenous and exogenous centrality. *Social Networks*, 32(4):339–344, 2010.

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

Attila Gursoy, Ozlem Keskin, and Ruth Nussinov. Topological properties of protein interaction networks from a structural perspective. *Biochemical Society Transactions*, 36(6):1398–1403, 2008.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th*

Python in Science Conference (SciPy2008), pages 11–15, Pasadena, CA USA, August 2008.

Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–806, 2005.

Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, Frederick P Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.

G Traver Hart, Arun K Ramani, Edward M Marcotte, et al. How complete are current yeast and human protein-interaction networks. *Genome Biol*, 7(11):120, 2006.

Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2(6):e88, 2006.

Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, 2005(2):96–103, 2005.

Ravi S Kamath, Andrew G Fraser, Yan Dong, Gino Poulin, Richard Durbin, Monica Gotta, Alexander Kanapin, Nathalie Le Bot, Sergio Moreno, Marc Sohrmann, et al. Systematic functional analysis of the caenorhabditis elegans genome using rnai. *Nature*, 421(6920):231–237, 2003.

Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In *Network analysis*, pages 16–61. Springer, 2005.

Harold J Leavitt. Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, 46(1):38, 1951.

Pierre Legrain and Luc Selig. Genome-wide protein interaction maps using two-hybrid systems. *FEBS letters*, 480(1):32–36, 2000.

Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC genomics*, 11(1):1, 2010.

Hao Luo, Yan Lin, Feng Gao, Chun-Ting Zhang, and Ren Zhang. Deg 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research*, page gkt1131, 2013.

Mitra Mirzarezaee, Babak N Araabi, and Mehdi Sadeghi. Features analysis for identification of date and party hubs in protein interaction network of saccharomyces cerevisiae. *BMC systems biology*, 4(1):1, 2010.

Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.

Tejaswini Narayanan, Merril Gersten, Shankar Subramaniam, and Ananth Grama. Modularity detection in protein-protein interaction networks. *BMC research notes*, 4(1):569, 2011.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, et al. The mips mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

Jose B Pereira-Leal, Anton J Enright, and Christos A Ouzounis. Detection of functional modules from protein interaction networks. *PROTEINS: Structure, Function, and Bioinformatics*, 54(1):49–57, 2004.

Jun Ren, Jianxin Wang, Min Li, Huan Wang, and Binbin Liu. Prediction of essential proteins by integration of ppi network topology and protein complexes information. In *Bioinformatics Research and Applications*, pages 12–24. Springer, 2011.

Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451, 2004.

Mihaela E Sardiu and Michael P Washburn. Building protein-protein interaction networks with proteomics and informatics tools. *Journal of Biological Chemistry*, 286(27):23645–23651, 2011.

William C Skarnes, Barry Rosen, Anthony P West, Manousos Koutsourakis, Wendy Bushell, Vivek Iyer, Alejandro O Mujica, Mark Thomas, Jennifer Harrow, Tony Cox, et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–342, 2011.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.

Ben Teng, Can Zhao, Xiaoqing Liu, and Zengyou He. Network inference from apms data: computational challenges and solutions. *Briefings in bioinformatics*, page bbu038, 2014.

Athanasios Typas and Victor Sourjik. Bacterial protein networks: properties and functions. *Nature Reviews Microbiology*, 13(9):559–572, 2015.

Alexander Veremyev, Oleg A Prokopyev, and Eduardo L Pasiliao. Critical nodes for distance-based connectivity and related problems in graphs. *Networks*, 66(3):170–195, 2015.

Chrysafis Vogiatzis, Alexander Veremyev, Eduardo L Pasiliao, and Panos M Pardalos. An integer programming approach for finding the most and the least central cliques. *Optimization Letters*, 9(4):615–633, 2015.

Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291, 2000.

Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, 2007.

Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: a molecular interaction database. *FEBS letters*, 513(1):135–140, 2002.

Ren Zhang and Yan Lin. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(suppl 1):D455–D458, 2009.

Ren Zhang, Hong-Yu Ou, and Chun-Ting Zhang. Deg: a database of essential genes. *Nucleic acids research*, 32(suppl 1):D271–D272, 2004.

Elena Zotenko, Julian Mestre, Dianne P O'Leary, and Teresa M Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8):e1000140, 2008.