Social Network Analysis and Mining (2023) 13:132
https://doi.org/10.1007/s13278-023-01135-3

ORIGINAL ARTICLE

The effect of the Katz parameter on node ranking, with a medical application

Hunter Rehm[1] · Mona Matar[2] · Puck Rombach[1] · Lauren McIntyre[2]

Received: 5 April 2023 / Revised: 18 July 2023 / Accepted: 11 September 2023 / Published online: 16 October 2023

Abstract
The Medical Extensible Dynamic Probabilistic Risk Assessment Tool (MEDPRAT), developed by NASA, is an event-based risk modeling tool that assesses human health and medical risk during space exploration missions. The Susceptibility Inference Network (SIN), a sub-element of MEDPRAT, is a prototype model informed with data that represents the probabilities of medical conditions progressing from one to another and the expected quality time lost associated with the disease progression for each condition. The work presented in this paper aims to determine which conditions in the SIN have the greatest effect on MEDPRAT-predicted medical risk. Here, we propose to measure this expected quality time lost using a weighted version of Katz centrality and investigate the effect of the α-parameter on the lengths of walks that significantly affect the ranking of nodes. To do this, we introduce a tool to compare different centrality measures in their node rankings. This general tool is of independent interest, as it considers that a relative ranking of two nodes by a centrality measure is unreliable if their scores are within a margin of error. In particular, we find an upper bound on the lengths of the walks that determine the node ranking up to this margin of error. If an application imposes a realistic bound on possible walk lengths, this set of tools may help determine a suitable value for α.

Keywords Network · Centrality · Katz parameter · Ranking

1 Introduction

Networks are structures that naturally appear in every aspect of life and are studied in a wide range of disciplines from sociology, biology, and engineering (Freeman 2004; Guze 2014; Newman 2018; Pavlopoulos et al. 2011). One common question in network theory is how to rank the nodes according to their importance, where importance can have many meanings depending on the application (Das et al. 2018). Many ranking algorithms are based on a, possibly weighted, count of walks in which a node is contained. Examples of such centrality measures are degree, betweenness (Freeman 1977), closeness (Bavelas 1950; Murray 1965; Freeman 1978), eigenvector (Bonacich 1972, 1987), PageRank (Page et al. 1999), the Estrada index (Estrada 2000), and Katz centrality (Katz 1953). We argue that the latter is suitable for our application, hence the focus of this paper.

We look at a model component developed by the National Aeronautics and Space Administration's (NASA) Human Research Program (HRP) called the Susceptibility Inference Network (SIN). This network represents the probability that simulated medical conditions may occur and progress to subsequent, clinically related conditions. When integrated with MEDPRAT, these relationships will produce results that are a more appropriate analog to the real-world medical system than the standing assumption that conditions are probabilistically independent. The SIN is currently a prototype, as the data used to inform it do not have the necessary credibility required by NASA standards 7150.2D and 7009A for use in decision support tools (https://nodis3.gsfc.nasa.gov/npg_img/N_PR_

✉ Hunter Rehm
hunter.rehm@nasa.gov

Mona Matar
mona.matar@nasa.gov

Puck Rombach
puck.rombach@uvm.edu

Lauren McIntyre
lauren.p.mcintyre@nasa.gov

[1] Department of Mathematics and Statistics, University of Vermont, Burlington, VT, USA

[2] NASA Glenn Research Center, Cleveland, OH, USA

7150_002D_/N_PR_7150_002D_.pdf, https://standards.nasa.gov/standard/nasa/nasa-std-3001-vol-1). Given the time and financial costs associated with evidence collection at this scale, there is significant motivation to focus the direction of those efforts toward conditions, or groups of conditions, whose relationships most influence medical risk outcomes.

Katz centrality, developed by Leo Katz in 1953 (Katz 1953), has been used in numerous applications (Fletcher and Wennekers 2018; Zhan et al. 2017). The Katz centrality of a node is a weighted count of all walks of any length starting at the node. Each walk of length k is weighted by α^k, where α is called the *Katz parameter*. We formally define Katz centrality in Sect. 1.1. Since the Katz parameter has a decaying effect, we can approximate the Katz centrality by ignoring the contribution of walks past a given length L. In Nathan and Bader (2017) and Nathan et al. (2017), the authors numerically explore this type of approximation. In Sect. 2, we give a lower bound on this value L (in terms of α) that guarantees a desired level of accuracy in terms of the Katz centrality and its node ranking.

This paper is organized as follows. Section 1.1 reviews useful graph theory concepts, definitions, and basic results. Here, we introduce the notion of ϵ-agreement of two centrality measures, which indicates their agreement regarding node rankings given an assumed ϵ margin of error in their node centrality scores. Section 1.2 introduces the SIN data set, which is the application of interest. In Sect. 2, we bound the error generated from approximating the Katz centrality by restricting the number of steps allowed in a walk and develop a relationship between that number and the Katz parameter α. An example of the relationship between α and the α-Katz centrality node ranking is given in Sect. 2 and our medical application in Sect. 3. Additionally, we assess the upper bound given in Sect. 2 to the true length in Sect. 4. Finally, the results and future work are addressed in Sect. 5.

1.1 Definitions

This section provides basic definitions for the graph-theoretical structures and tools used for the results in Sect. 2.

Definition 1 (*Weighted, directed network*) Let $N = (V, E, w)$ be an *edge-weighted, directed network* consisting of V, the set of n *nodes*, $E \subseteq V \times V$, the set of *edges*, and a weight function $w : E \to \mathbb{R}^+$.

We represent such a network by an *adjacency matrix* $A = A(N)$, where the entry A_{ij} is the weight of the edge from node i to node j, or $A_{ij} = 0$ if there is no edge from i to j in N. Let W be an n-dimensional vector of non-negative node weights. In a setting where edges and/or nodes are unweighted, weights in A and W are set to 1.

For example, in our application in Sect. 3, our weighted, directed network has nodes that represent medical conditions, and the node weights represent their severity, while edge weights represent the probability of one medical condition progressing to another.

The *spectral radius* ρ of N (or A) is the maximum modulus of the eigenvalues of A. A *walk of length* k from node u to v is a sequence of k edges $(v_i, v_{i+1}) \in E, i \in [1, k]$ such that $v_1 = u$ and $v_{k+1} = v$. The *distance* from u to v is the length of a shortest walk from u to v. The *k-hop neighborhood* of a node $v \in V$ is the set of nodes at a distance less than or equal to k from v. A *centrality measure* is a function that assigns a real number to each node, to evaluate its relative importance to other nodes. Each centrality measure gives a (partial) *ranking* of the nodes, which reflects their relative importance.

Our focus is on Katz centrality, a parameterized centrality measure whose parameter α takes in walk length considerations.

Definition 2 (*Katz centrality*) Let N be an edge-weighted, directed network with node weights W. Let $A = A(N)$ with spectral radius ρ, and let $\alpha \in (0, 1/\rho)$. The *α-Katz centrality* vector (De la Cruz Cabrera et al. 2019; Estrada and Higham 2010; Katz 1953) is defined as

$$C(\alpha) = \left(\sum_{k=1}^{\infty} \alpha^k A^k \right) \cdot W = \left((I - \alpha A)^{-1} - I \right) \cdot W.$$

The *α-Katz score* of a particular node i can then be expressed as

$$C(\alpha)_i = \sum_{k=1}^{\infty} \sum_{j=1}^{n} W_j \alpha^k \left(A^k\right)_{ij}.$$

The *(α, ℓ)-Katz centrality* vector (Acar et al. 2009; Béres et al. 2018; Lu et al. 2010) is

$$C(\alpha, \ell) = \left(\sum_{k=1}^{\ell} \alpha^k A^k \right) \cdot W$$

and for a particular node i, the *(α, ℓ)-Katz score* can be written as

$$C(\alpha, \ell)_i = \sum_{k=1}^{\ell} \sum_{j=1}^{n} W_j \alpha^k \left(A^k\right)_{ij}.$$

Both $C(\alpha)$ and $C(\alpha, \ell)$ measure the *downstream* influence of nodes since they are weighted sums over outgoing walks. Replacing the matrix A by its transpose A^T reverses edge directions, taking weighted sums over incoming walks instead and measuring the *upstream* influence (De la Cruz Cabrera et al. 2019; Newman 2018).

Definition 3 provides a tool to compare centrality measures purely in terms of their relative node rankings. Intuitively, we may set a threshold ϵ for a centrality measure C, such that $|C_i - C_j| \geq \epsilon$ implies that C provides a relative ranking of nodes i and j. If $|C_i - C_j| < \epsilon$, we cannot reliably recover a ranking from C. For two centrality measures C and C', we compare their rankings and conclude that they agree on a ranking if they agree for every node pair where both rankings are reliable.

Definition 3 (ϵ*-agreement*) Let N be a weighted, directed network, $\epsilon, \epsilon' > 0$, and C and C' be centrality measures. The nodes $i, j \in V(N)$ are (ϵ, ϵ')*-properly ranked with respect to* C *and* C' if the following holds:

1. $|C_i - C_j| < \epsilon$ or $|C'_i - C'_j| < \epsilon'$,
2. otherwise, $C_i - C_j$ and $C'_i - C'_j$ have the same sign.

We say that C and C' (ϵ, ϵ')*-agree on* N if every pair of nodes in N is (ϵ, ϵ')-properly ranked with respect to C and C'. If $\epsilon = \epsilon'$, we simply say ϵ-proper ranking and ϵ-agreement.

Definition 4 Let N be a weighted, directed network, $\epsilon > 0$, $\alpha \in (0, 1/\rho)$. We let

$$L_{\alpha,\epsilon}(N) = \min\{\ell \mid C(\alpha) \text{ and } C(\alpha, \ell)\ \epsilon\text{-agree on } N\}.$$

When the parameters are clear from the context, we will let $L = L_{\alpha,\epsilon}(N)$.

Proposition 1 *Let* C *and* C' *be two centrality measures on a network* N. *If*

$$\|C - C'\|_\infty < \epsilon,$$

then C *and* C' ϵ*-agree.*

Proof Suppose for the sake of contradiction that C and C' do not ϵ-agree. Then, there exists a pair of nodes $u, v \in V(N)$ that is not ϵ-properly ranked. By part (1) of Definition 3, we have $|C_u - C_v| > \epsilon$ and $|C'_u - C'_v| > \epsilon$. By part (2), without loss of generality, we have $C_u - C_v < 0$ and $C'_u - C'_v > 0$.

We have

$$C'_u - C_u = \underbrace{C'_u - C'_v}_{>\epsilon} + \underbrace{C'_v - C_v}_{>-\epsilon} + \underbrace{C_v - C_u}_{>\epsilon} > \epsilon,$$

which contradicts that $\|C - C'\|_\infty < \epsilon$. □

Proposition 2 *Let* C *and* C' *be two centrality measures on a network* N *such that for all* $v \in V(N)$, $0 \leq C_v - C'_v < 2\epsilon$, *then* C *and* C' ϵ*-agree.*

Proof Suppose for the sake of contradiction that C and C' do not ϵ-agree. Then, there exists a pair of nodes $u, v \in V(N)$ that is not ϵ-properly ranked. By part (1) of Definition 3, we have $|C_u - C_v| > \epsilon$ and $|C'_u - C'_v| > \epsilon$. By part (2), without loss of generality, we have $C_u - C_v < 0$ and $C'_u - C'_v > 0$.

We have

$$C_u - C'_u = \underbrace{C_u - C_v}_{<-\epsilon} + \underbrace{C_v - C'_v}_{<2\epsilon} + \underbrace{C'_v - C'_u}_{<-\epsilon} < 0,$$

which contradicts that $C_u - C'_u \geq 0$. □

In Sect. 2, we use this notion to compare two closely related centrality measures, $C(\alpha)$ and $C(\alpha, \ell)$, and we therefore only use ϵ-proper ranking and ϵ-agreement. However, we state the definition here in a more general form. It can be used to compare any pair of centrality measures, even if their distributions of values differ significantly. The vector $C(\alpha, \ell)$ converges to $C(\alpha)$ as $\ell \to \infty$. In Theorem 1, we show that this implies that for all $\epsilon > 0$, there exists an L so that for any $\ell > L$, $C(\alpha)$ and $C(\alpha, \ell)$ ϵ-agree.

1.2 Susceptibility Inference Network

The Medical Extensible Dynamic Probabilistic Risk Assessment Tool (MEDPRAT) developed by NASA is an event-based risk modeling tool that assesses human health and medical risk during space exploration missions (McIntyre et al. 2020, 2022). One of its key features is the ability to represent and simulate the relationships between medical events. The Susceptibility Inference Network (SIN) captures these relationships in an internal data structure.

The SIN is a directed network where nodes represent medical conditions. The data in this network are subject matter expert informed and are currently a prototype. There is an edge from u to v if medical condition u can progress into medical condition v. This directed edge (u, v) is weighted by the probability that such a progression occurs. Note that a medical condition may progress to multiple other conditions simultaneously or to no other conditions. Therefore, this matrix is not a transition matrix. The SIN currently has 99 nodes and 1078 edges. Medical conditions included are, for example, acute radiation syndrome, which has many outgoing edges toward other medical conditions. On the other hand, anxiety has many incoming edges.

This expert-informed data do not contain information about the time it takes for progressions to occur. As a simplified model, we view the SIN as a Dynamic Bayesian Network (Dagum et al. 1992). Each node in the SIN has an associated weight that evaluates the severity of the condition regardless of the progression from or to that condition. This severity of a condition is quantified by Quality Time Lost (QTL), which we call *primary QTL* of that condition. The primary QTL measures the productive time a crew member is expected to lose and is equal to the ith entry W_i of the weight vector W. Primary QTL is a measure in days of the time astronauts cannot perform tasks due to being afflicted by medical conditions and is one of the model outputs from MEDPRAT; the data set informing the model is an evidence-based collection of condition incidence and outcome data.

In our condition progression networks, each edge (i,j) is weighted by A_{ij}, which is the probability that condition j is present at time $t+1$ given that condition i is present at time t. Then, this edge contributes $A_{ij}W_j$ to $\mathbb{E}[\,\mathrm{QTL}_i]$. We assume that progressions occur independently, and therefore a path of length two from i to j via a node k contributes $A_{ik}A_{kj}W_j$ to $\mathbb{E}[\,\mathrm{QTL}_i]$. Under the assumption of uninterrupted progressions, we have, in general,

$$\mathbb{E}[\,\mathrm{QTL}_i] = \sum_{k=0}^{\infty}\sum_{j=1}^{n} W_j\left(A^k\right)_{ij} = W_i + C(1)_i.$$

Note that W_i is the primary QTL of node i and $C(1)_i$ the subsequent QTL under the assumption of uninterrupted progression. We highlight a few subtleties in this model. Note that we allow two types of cycles in our network: There may be multiple directed paths from a condition i to a condition j, in which case j contributes multiple times to the total expected QTL of i. There may also be directed cycles, wherein a condition i contributes multiple ways to its total expected QTL. We motivate this in terms of the application later. First, we consider an illustration. Table 1 contains examples of small condition progression networks.

To make this estimate more realistic, we consider that the progression of conditions will likely be interrupted by medical interventions and time constraints on the mission. The parameter α provides a damping factor that decreases the weight of walks as they get longer. This application, therefore, illustrates the importance of using realistic walk lengths to guide the choice of α. We provide a theoretic foundation for this in Sect. 2. In Sect. 3, we discuss how different values of α produce different rankings for the SIN due to subsequent QTL.

2 The Katz parameter and walk length

This section describes the relationship between maximum walk lengths ℓ and the parameter α. In Theorem 1, we find a lower bound on ℓ that guarantees that $C(\alpha)$ and $C(\alpha,\ell)$ ϵ-agree. This sheds light on the length of walks that decide the ranking provided by $C(\alpha)$. We provide a small, illustrative example of the effect of α on node rankings.

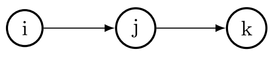
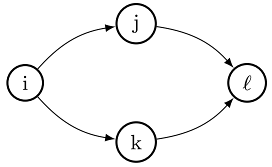
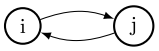
Lemma 1 gives an upper bound on the difference between values in $C(\alpha)$ and $C(\alpha,\ell)$.

Lemma 1 (Absolute Error Tolerance) *Let $p \in \{1,2,\infty\}$ and $\alpha \in (0,1/\rho)$. Then*

$$\|C(\alpha) - C(\alpha,\ell)\|_{\infty} \leq \left(\alpha\|A\|_p\right)^{\ell}\|C(\alpha)\|_p := \epsilon_{\ell}.$$

Proof First, note that $\|V\|_{\infty} \leq \|V\|_2 \leq \|V\|_1$ for all vectors V. Furthermore, the p-norm is sub-multiplicative. We have

Table 1 Three networks with the expected QTL of node i

Condition Progression Network	$\mathbb{E}(\mathrm{QTL}_i)$
	$W_i + A_{ij}W_j + A_{ij}A_{jk}W_k$
	$W_i + A_{ij}W_j + A_{ik}W_k + (A_{ij}A_{j\ell} + A_{ik}A_{k\ell})W_\ell$
	$\sum_{k=0}^{\infty}(A_{ij}A_{ji})^k W_i + A_{ij}\sum_{k=0}^{\infty}(A_{ij}A_{ji})^k W_j = \frac{W_i + A_{ij}W_j}{1 - A_{ij}A_{ji}}$

$$\begin{aligned}\|C(\alpha)-C(\alpha,\ell)\|_\infty &\le \|C(\alpha)-C(\alpha,\ell)\|_p\\ &= \left\|\left(\sum_{k=\ell+1}^{\infty}\alpha^k A^k\right)\cdot W\right\|_p\\ &= \left\|\alpha^\ell A^\ell\left(\sum_{k=1}^{\infty}\alpha^k A^k\right)\cdot W\right\|_p\\ &\le \alpha^\ell \|A^\ell\|_p\left\|\left(\sum_{k=1}^{\infty}\alpha^k A^k\right)\cdot W\right\|_p\\ &\le \left(\alpha\|A\|_p\right)^\ell \|C(\alpha)\|_p.\end{aligned}$$

□

In Lemma 2, we show that there exists an L so that for all $\ell > L$, the difference between the scores in $C(\alpha)$ and $C(\alpha,\ell)$ is small.

Lemma 2 (Error Tolerance Guarantee) *Let* $p \in \{1,2,\infty\}$, $\alpha \in (0, 1/\|A\|_p)$ *and* $\epsilon > 0$. *If*

$$\ell > \log_{\alpha\|A\|_p}\left(\frac{2\epsilon}{\|C(\alpha)\|_p}\right) := L_{up}$$

then $\|C(\alpha)-C(\alpha,\ell)\|_\infty < 2\epsilon$.

Proof Note that $\alpha < 1/\|A\|_p \le 1/\rho$. By Lemma 1, it suffices to show that $\epsilon_\ell < 2\epsilon$ when $\ell > L_{up}$. We have

$$\begin{aligned}\epsilon_\ell &= \|C(\alpha)\|_p\left(\alpha\|A\|_p\right)^\ell\\ &< \|C(\alpha)\|_p\left(\alpha\|A\|_p\right)^{L_{up}}\\ &= \|C(\alpha)\|_p\left(\alpha\|A\|_p\right)^{\log_{\alpha\|A\|_p}\left(\frac{2\epsilon}{\|C(\alpha)\|_p}\right)}\\ &= \|C(\alpha)\|_p\frac{2\epsilon}{\|C(\alpha)\|_p}\\ &= 2\epsilon.\end{aligned}$$

□

Lemma 2 bounds the difference between $C(\alpha)$ and $C(\alpha,\ell)$ when ℓ is large enough. Theorem 1 shows that this also ensures that $C(\alpha)$ and $C(\alpha,\ell)$ ϵ-agree.

Theorem 1 *Let* $p \in \{1,2,\infty\}$, $\alpha \in (0, 1/\|A\|_p)$, $\epsilon > 0$ *and* L_{up} *be as in Lemma 2. If* $\ell > L_{up}$ *then* $C(\alpha)$ *and* $C(\alpha,\ell)$ ϵ*-agree.*

Proof By Lemma 2 we have that $\|C(\alpha)-C(\alpha,\ell)\|_\infty < 2\epsilon$. Therefore, by Proposition 2, we have that $C(\alpha)$ and $C(\alpha,\ell)$ ϵ-agree. □

We choose ϵ so that two nodes with Katz scores within ϵ of each other can be considered equivalent in terms of ranking. Of course, such a value must be chosen relative to the Katz scores. In Corollary 1, we suggest letting ϵ be some fraction of $\|C(\alpha)\|_p$, and α a fraction of $1/\|A\|_p$ for $p \in \{1,2,\infty\}$.

Corollary 1 (Relative Error) *Let* $p \in \{1,2,\infty\}$ *and* $\alpha_0, \epsilon_0 \in (0,1)$. *For* $\alpha = \alpha_0/\|A\|_p$, *and* $\epsilon = \epsilon_0\|C(\alpha)\|_p$, *if* $\ell > \log_{\alpha_0}(\epsilon_0) = L_{up}$ *then* $C(\alpha)$ *and* $C(\alpha,\ell)$ ϵ*-agree.*

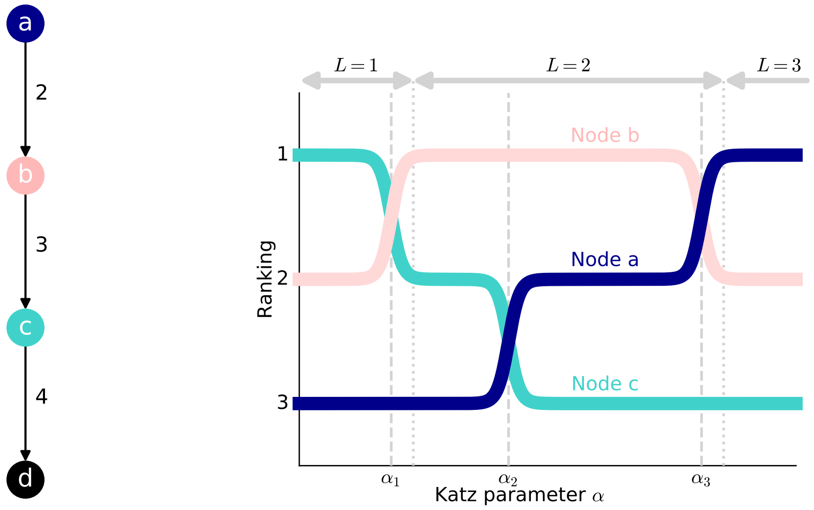
We present a small example illustrating the effect of α on node rankings. Consider the edge-weighted directed network N in Fig. 1a. Let α_1 be the positive solution to $C(\alpha)_b = C(\alpha)_c$, let α_2 be the positive solution to $C(\alpha)_a = C(\alpha)_c$, and let α_3 be the positive solution to $C(\alpha)_a = C(\alpha)_b$. When α is small, the walks of length 1 determine the ranking. As α increases, walks of length 2 and later walks of length 3 become more important. Node c has the most walks of length 1, node b the most of length 2, and node a the most of length 3, and they are each ranked on top for different ranges of α.

Figure 1b also shows the value of $L = L_{\alpha,\epsilon}(N)$ as a function of α. For example, at α_1, the walks of length 2 become significant enough for node b to overtake node c in the ranking. Therefore, L switches from 1 to 2. This switch happens at a value of α slightly greater than α_1, once the difference in scores of b and c has exceeded ϵ. Note that the ranking switch of node a and c at α_2 is also due to paths of length 2 gaining significance and does not cause a change in L.

3 Application to the Susceptibility Inference Network

We use α-Katz centrality to rank medical conditions in the SIN by their expected subsequent QTL. We run a systematic sensitivity analysis on α by comparing the top 10 conditions given by α-Katz centrality for $\alpha \in (0, 1.75)$ in Fig. 2. We inspect the distribution of the differences between Katz scores of nodes and set ϵ so that 5% of the differences are less than the chosen ϵ. This implies that at least 95% of the pairs are ranked correctly. The values $\alpha = 1$ and $\alpha = 0.4$ are of particular interest as the α- and (α,ℓ)-Katz centrality ϵ-agree at realistic condition progression lengths. For $\alpha = 1$, (α,ℓ)-Katz centrality and α-Katz centrality ϵ-agree when $\ell \ge 5$, and for $\alpha = 0.4$, (α,ℓ)-Katz centrality and α-Katz centrality ϵ-agree when $\ell \ge 3$. For more on how to choose α, ϵ, and ℓ to suit a particular application, please see Sect. 5. Figure 4 illustrates the subnetwork of the SIN containing the 12 conditions that appear in Fig. 2. The thickness of each arrow pointing outward in Fig. 4 represents the sum of the edges weights leaving that node.

Fig. 1 The relationship between the Katz parameter α and the node ranking in a directed, edge-weighted example network



(a) An example of a weighted directed network.

(b) The node rankings from the Katz scores of the network in Figure 1a.

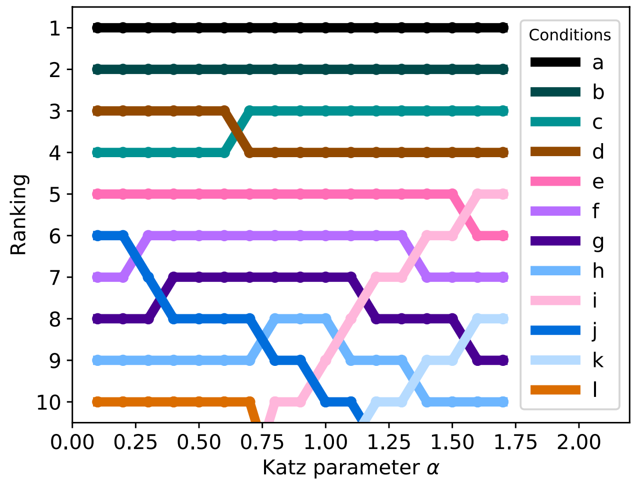


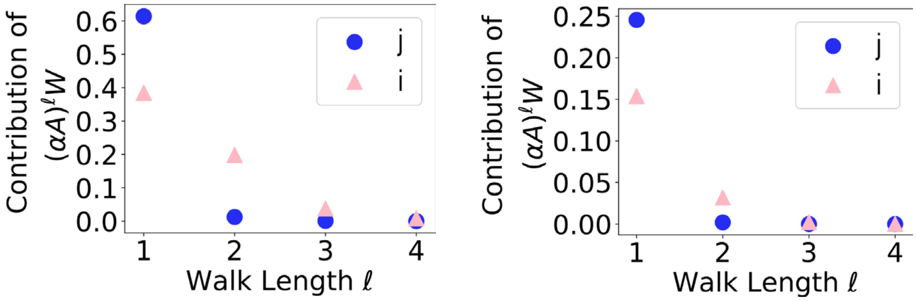
Fig. 2 The effect of the Katz parameter α on the ranking of α-Katz centrality in the Susceptibility Inference Network (SIN)

We note that one significant change in the ranking as α decreases from 1 to 0.4 is the one that swaps the positions of nodes i and j. Figure 3 plots the contributions of walks of length ℓ to the Katz scores of these two nodes. When $\alpha = 1$, the contribution of walks of lengths 2 and 3 contributes enough to the score of node i to place it above j. When $\alpha = 0.4$ these longer walks contribute significantly less, lowering the ranking of node i to fall below that of j. There is no universal optimal value of α. The choice of α should depend on path lengths considered most important. For example, in our application, length of the missions and available medical interventions affect how realistic any number of progressions of medical conditions is.

4 Testing the upper bound on simulated data

We would like to better understand for which values of ℓ that $C(\alpha)$ and $C(\alpha, \ell)$ ϵ-agree. Theorem 1 gives an upper bound on $L_{\alpha,\epsilon}(N)$. In Fig. 5, we compare the upper bound L_{up} to $L_{\alpha,\epsilon}(N)$ on two families of undirected graphs. At each instance of α_0 in Fig. 5a, b, we sample 10 graphs from each

Fig. 3 Contribution of walks of length ℓ to the scores of nodes i and j in 1-Katz centrality (**a**) and 0.4-Katz centrality (**b**). The centrality of j relies on short paths and i on long paths, and the specific length depends on which value of α is used



(a) $\alpha = 1$

(b) $\alpha = 0.4$

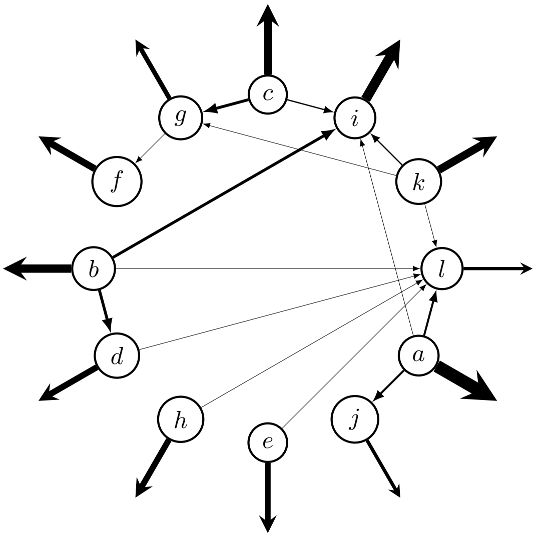


Fig. 4 Subnetwork of the SIN with the 12 most influential conditions and the weighted edges connecting them. The thickness of the edges that point outward illustrates the total outgoing edge weight toward nodes in the remainder of the network

specified graph family and plot the average upper bound and average $L_{\alpha,\epsilon}(N)$ across the samples.

In Fig. 5a, we sample from the Erdős–Rényi model $G(n, p)$ where $n = 1000$ and $p = 0.008$. In Fig. 5b we sample from the Chung-Lu model, which takes a list of node degrees as input. To create this list, we individually sample 1000 numbers from the negative binomial distribution. This distribution takes a probability p of success and a number n of desired successes. We use $p = 0.1$ as the probability of success and $n = 3$ as the number of desired successes.

The Chung–Lu model, as described here, produces graphs with a longer-tailed degree distribution than graphs sampled from the Erdős–Rényi model. In Fig. 5, the value for ϵ that we use is calculated using the same technique as in Sect. 3 for each iteration. We create a list of pairwise differences between the Katz scores and set ϵ so that 5% of the differences are less than it.

5 Discussion and conclusions

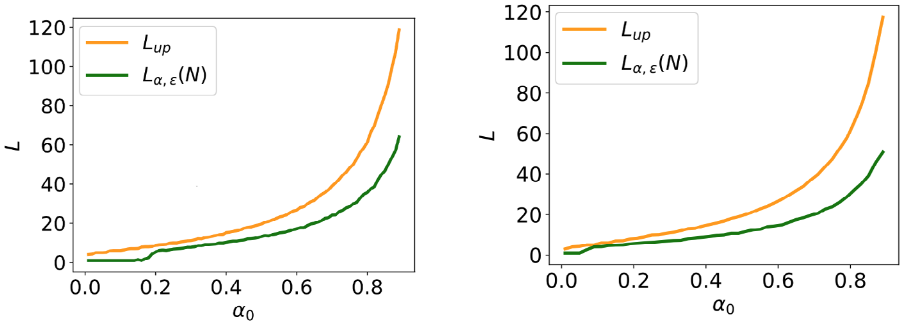
This paper introduces a tool to help compare different centrality measures. We apply this to α- and (α, ℓ)-Katz centrality to help better understand the effect of the α-parameter on the walk lengths considered when it comes to the ranking of nodes. For a given α, we provide an upper bound on the walk length L so that when $\ell > L$, the Katz scores in α- and (α, ℓ)-Katz centrality are within 2ϵ of each other. We show that two nodes with both centrality measures differing by at least ϵ are in the same order in both rankings when $\ell > L$.

The goal is to find a minimal value of L such that α-Katz and (α, ℓ)-Katz centrality ϵ-agree for all $\ell \geq L$, and Theorem 1 provides an upper bound. The value ϵ reflects the precision of the centrality scores in a given application. If two values are within ϵ of each other, they are not relatively ranked reliably and should be considered equivalent. The choice of ϵ should therefore depend on the application and its distribution of scores. We inspect the distribution of the differences between Katz scores of nodes and set ϵ so that 5% of the differences are less than the chosen ϵ. Theorem 1 can guide the choice of α, by relating, together with ϵ, to the maximum walk lengths L of interest. One can find a stronger upper bound by iteratively increasing L until for all $\ell \geq L$, $\|C(\alpha) - C(\alpha, \ell)\| < 2\epsilon$. This guarantees ϵ-agreement, although ϵ-agreement may still happen sooner, so care should be taken when interpreting these bounds.

All of the results in this paper require α to be in $(0, 1/\|A\|_2)$, a subset of the possible values that can be used as a Katz parameter. It may be possible to extend these results to all Katz parameters, namely, values in $(0, 1/\rho)$. These ranges match for undirected graphs, so this extension only applies to directed, edge-weighted graphs. The analysis done in this paper might be applied in a similar way to address upstream and downstream influence together, as well as other centrality measures such as eigenvector centrality and the Estrada index.

We show the effect of changing the α parameter in the Susceptibility Inference Network. In this case, changes in

Fig. 5 Comparing the upper bound to $L_{\alpha,\epsilon}(N)$ using the Erdős–Rényi model $G(1000, 0.008)$ in (**a**) and the Chung-Lu model with degrees sampled from the negative binomial distribution in (**b**)



(a) Erdős-Rényi.

(b) Chung-Lu.

the ranking were visible even among the top 10 nodes, and they may have significant implications for decision-making. It is, therefore, important that the choice of α is made carefully and tailored to each application. This also holds for ϵ.

The work presented in this paper will be leveraged to identify subsets of the prototype SIN that are expected to produce the largest effect in MEDPRAT and therefore to act as a scalable road map for future work, which will be focused on collecting and validating higher credibility clinical evidence. Given the significant cost associated with evidence collection, it is critical to narrow the scope and focus the effort where it will be most valuable, which is to say where interactions between conditions produce the largest change in spaceflight medical risk.

Acknowledgements The authors would like to acknowledge the Cross-Cutting Computational Modeling Project team at Glenn for their expertise, especially Dr. Drayton Munster, for the helpful discussions and feedback throughout this work. The authors thank NASA's Human Research Program, the University Space Research Association (USRA), and HX5 for supporting this work. Hunter Rehm was partially supported by the Vermont Space Grant Consortium (VTSGC) through the Graduate Fellowship Program.

Author contributions LM compiled the SIN matrix. HR and MM researched the background and conceptualized the research direction. HR derived the formulas and generated all the figures in the manuscript. HR and MM wrote the first draft of the manuscript. HR and PR updated the presentation of the proofs and results. All authors engaged in technical discussions, commented on previous versions of the manuscript and approved the final manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

Acar E, Dunlavy DM, Kolda TG (2009) Link prediction on evolving data using matrix and tensor factorizations. In: 2009 IEEE international conference on data mining workshops. IEEE, pp 262–269

Bavelas A (1950) Communication patterns in task-oriented groups. J Acoust Soc Am 22(6):725–730

Béres F, Pálovics R, Oláh A, Benczúr AA (2018) Temporal walk based centrality metric for graph streams. Appl Netw Sci 3(1):1–26

Bonacich P (1972) Technique for analyzing overlapping memberships. Sociol Methodol 4:176–185

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92(5):1170–1182

Dagum P, Galper A, Horvitz E (1992) Dynamic network models for forecasting. Uncertain Artif Intell 8(1):41–48

Das K, Samanta S, Pal M (2018) Study on centrality measures in social networks: a survey. Soc Netw Anal Min 25:1–11

De la Cruz Cabrera O, Matar M, Reichel L (2019) Analysis of directed networks via the matrix exponential. J Comput Appl Math 355:182–192

Estrada E (2000) Characterization of 3D molecular structure. Chem Phys Lett 319(5–6):713–718

Estrada E, Higham DJ (2010) Network properties revealed through matrix functions. SIAM Rev 52(4):696–714

Fletcher JM, Wennekers T (2018) From structure to activity: Using centrality measures to predict neuronal activity. Int J Neural Syst 28(02):1750013

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 56:35–41

Freeman L (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

Freeman L (2004) The development of social network analysis. Study Sociol Sci 1(687):159–167

Guze S (2014) Graph theory approach to transportation systems design and optimization. TransNav Int J Mar Navig Saf Sea Transp 8:57–62

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Lu Z, Savas B, Tang W, Dhillon IS (2010) Supervised link prediction using multiple sources. In: 2010 IEEE international conference on data mining. IEEE, pp 923–928

McIntyre L, Leinweber L, Myers JG (2020) Dynamic medical risk assessment supported by inference networks. In: Human research program investigators workshop (HRP IWS 2020), number GRC-E-DAA-TN77298

McIntyre L, Myers JG, Leinweber L, Prelich M, Gasiewski C, Lovell M, Prabhu R (2022) A model based approach to estimating human spaceflight medical risk. In: Committee on space research (COSPAR)

Murray A (1965) Beauchamp. An improved index of centrality. Behav Sci 10(2):161–163

NASA. NASA software engineering requirements (NPR 7150.2D). Office of the Chief Engineer. https://nodis3.gsfc.nasa.gov/npg_img/N_PR_7150_002D_/N_PR_7150_002D_.pdf

NASA. NASA space flight human-system standard volume 1, revision a: crew health. NASA Technical Standard. https://standards.nasa.gov/standard/nasa/nasa-std-3001-vol-1

Nathan E, Sanders G, Fairbanks J, Henson VE, Bader DA (2017) Graph ranking guarantees for numerical approximations to Katz centrality. Procedia Comput Sci 108:68–78

Nathan E, Bader DA (2017) Approximating personalized Katz centrality in dynamic graphs. In: International conference on parallel processing and applied mathematics. Springer, pp 290–302

Newman M (2018) Networks. Oxford University Press, Oxford

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab

Pavlopoulos GA, Maria S, Moschopoulos CN, Soldatos TG, Sophia K, Jan A, Reinhard S, Bagos PG (2011) Using graph theory to analyze biological networks. BioData Min 4(1):1–27

Zhan J, Gurung S, Parsa SPK (2017) Identification of top-*k* nodes in large networks using Katz centrality. J Big Data 4(1):1–19