Option 2: Katz and NBTW Centrality Measures: Comparison and Impact of Measurement Errors

[a]Candidate Number: 1092519, Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

In this report, we compare two walk-based centralities: Katz and Non-Backtracking Walk (NBTW). We first explain the motivation and mathematical foundations behind these centrality measures, defining them and presenting results related to their radii of convergence and limiting behaviour. We then compare them theoretically and empirically using real-world data of Twitter reciprocated mentions. We analyse the impact of the attenuation factor on their rankings and the phenomenon of localisation. We also propose a new, intuitive way of comparing localisation effects based on the allocation of the 50% of the total weight from the centrality measure. In the final part of the report, we examine the impact of false positive and false negative edges on these centrality measures. To evaluate the degradation of centrality measures, we use statistical methods proposed in various research papers for assessing robustness. Despite differences between Katz and NBTW centralities, their robustness to network errors is strikingly similar. We conclude that false positives have less impact on the rankings of top nodes than false negatives.

Katz centrality | NBTW centrality | Network error

Centrality measures are crucial in network analysis to understand complex systems. In simple terms, centrality measures quantify the importance of each node within a network. Centrality measures are widely applied in social sciences (1, 2) and biological settings (3, 4) to identify the most dominant nodes, such as vital genes or proteins. The simplest example of a centrality measure is the degree centrality, assessing a node's importance based on its number of links.

While degree centrality offers a direct measure based on connections, it overlooks the significance of a node's connections. Katz (5) addressed this by introducing a walk-based centrality that relies on counting possible walks from a node while penalising (by some attenuation factor) contributions from lengthy walks. This approach is both intuitive and computationally convenient since we are dealing with walks, not paths, allowing for the analysis of large-scale networks. Although Katz centrality provides valuable insights, it factors in walks that immediately return to the previously visited node, which intuitively do not convey much importance. To refine the analysis, we explore Non-Backtracking Walk (NBTW) centrality, which is a generalised version of Katz centrality that excludes these types of walks, as outlined in (6–8).

Interpreting centrality measures applied on real-world networks requires caution due to the susceptibility of these networks to measurement errors. For example, Brewer (9) found that, when surveyed, individuals forget an average of 20% of their friends. Wang (10) categorized measurement errors into six groups, including false-positive nodes and edges, false-negative nodes and edges, and false aggregation and disaggregation. These errors can significantly impact centrality measurements, as noted in (11–14). In this report, we analyse the impact of false-positive and false-negative edges on Katz and NBTW centralities.

This report is divided into three primary sections: summarising the mathematical theory behind Katz and NBTW centralities, comparing these centralities on the real-world data (15), and investigating their robustness in the presence of false-positive and false-negative edges.

1. Mathematical framework

Key definitions. In simple terms, **network** is a collection of nodes, some of which are linked. Mathematically, a network is defined as a graph $G = (V, E)$, where V represents the set of nodes (vertices), and $E \subseteq V \times V$ is the set of links (edges). In general, edges can be either directed (pointing from one vertex to another) or weighted (e.g. with weights representing distances between nodes). However, in this report, our focus will be on *unweighted undirected networks*. It's worth noting that the results discussed in this report can be extended to directed networks as in (16) and weighted directed networks as in (17).

We enumerate edges with numbers from 1 to $n = |V|$ and use an adjacency matrix $A = (a_{ij})$ to represent a network algebraically. A is a binary $n \times n$ matrix such that $a_{ij} = 1$ if there exists an edge between nodes $i, j \in V$; otherwise, $a_{ij} = 0$. We do not allow self-loops, i.e. $a_{ii} = 0$ for all $i \in V$. For undirected graphs, the adjacency matrix A is symmetric, i.e. $a_{ij} = a_{ji}$ for all $i, j \in V$. Unless specified otherwise, we choose to employ the above notations and conventions throughout the remainder of the report when referring to networks.

In this report, we will work with walk-based centralities. A sequence $v_0 v_1 \dots v_k$ is a **walk** in a network G if $v_0, v_1, \dots, v_k$ are nodes (not necessarily distinct) such that $v_i v_{i+1} \in E$ for each $i = 0, 1, \dots, k-1$. The **length** of the walk is the number of edges it contains, here t. In particular, a single node is a walk of length 0.

Significance Statement

Determining the importance of nodes is crucial in various fields, such as identifying influential individuals in social networks or essential proteins in biology. Centrality measures, specifically walk-based centralities like Katz and NBTW centrality measures, provide an easy way of assessing a node's global importance. However, networks may involve errors, such as spurious edges or omitted connections, stemming from data collection or processing. It is crucial to assess how these errors affect centrality measure rankings to ensure accurate and reliable network analyses. It is vital in fields where precise node classification is necessary for meaningful insights, such as managing an epidemic network.

Furthermore, a **centrality measure** c is a real-valued function that assigns values to all nodes in a graph. Importantly, it remains invariant to structure-preserving mappings, meaning that centrality values depend solely on the inherent structure of the graph.

Katz centrality.

Motivation. Suppose we want to quantify the significance of a given node i based on the number of walks originating from that node. However, we encounter two obstacles with this approach:

(i) Counting all the walks originating from node i would be inviable because there are infinitely many of them. Indeed, one can perpetually traverse back and forth between the initial node and any of its neighbours.

(ii) Longer walks are less informative about the node's significance than shorter ones.

Katz centrality addresses these issues by scaling the importance of each walk of length k by a factor of α^k, where α is an attenuation factor. α should be between 0 to 1, but a more rigorous discussion will follow in the next subsection.

General formula and radius of convergence. As demonstrated in (18), the (i,j)-th entry of the k-th power of the adjacency matrix, $\left(A^k\right)_{ij}$, is the number of walks of length k from node i to node j. Consequently, the (i,j)-th entry of the Katz centrality matrix:

$$C_{\text{Katz}}(A,\alpha) = I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \dots \qquad [1]$$

is the weighted sum of walks from node i to node j. Katz centrality of node i can be defined as follows:

$$\text{Katz}_i(A,\alpha) = \sum_{j=1}^{n} \left(C_{\text{Katz}}(A,\alpha)\right)_{ij}.$$

In matrix theory, it is well-known that the Neumann series in equation 1 converges to $(I-\alpha A)^{-1}$ for all (possibly complex) values of α with a modulus less than $1/\rho(A)$. Here, $\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$ denotes the **spectral radius** of A. Therefore, as $\text{Katz}_i(A,\alpha)$ is the sum of the ith row of the Katz centrality matrix, if $\mathbf{c}_{\text{Katz}}$ is a vector with the i-th entry equal to $\text{Katz}_i(A,\alpha)$, it satisfies the linear system $(I-\alpha A)\mathbf{v}_{\text{Katz}} = \mathbf{1}$.

Assuming the network is connected, A is irreducible (meaning that for each pair of nodes i,j, there is a natural number k such that $\left(A^k\right)_{ij} > 0$). According to the Perron-Frobenius Theorem (19) for irreducible nonnegative matrices, $\rho(A)$ is also the largest positive eigenvalue of A, called the **Perron-Frobenius eigenvalue**. If the network is not connected, $\rho(A)$ is still an eigenvalue of A but might be 0.

To maintain interpretability, we require $0 < \alpha < 1/\rho(A)$. These bounds ensure the convergence of the power series in 1 to $(I-\alpha A)^{-1}$. Moreover, they imply that $I-\alpha A$ is a nonsingular M-matrix—an invertible matrix with all off-diagonal entries less than or equal to zero and with eigenvalues whose real parts are nonnegative. As a result, by (20), $(I-\alpha A)^{-1}$ is nonnegative so that we can use the row sums for ranking of nodes.

Limiting behaviour. Under the assumption of a connected network, according to Corollary 1 in (21), the rankings produced by Katz centrality converge to:

(i) Degree centrality ranking when $\alpha, \to 0+$;

(ii) Eigenvalue centrality ranking ($A\mathbf{v}_{\text{eigen.}} = \lambda\mathbf{v}_{\text{eigen.}}$, where λ is the Perron-Frobenius eigenvalue; for reference (22)) when $\alpha \to \frac{1}{\rho(A)}-$.

NBTW centrality.

Motivation. As discussed earlier, Katz centrality considers walks that may be less impactful. Specifically, for every edge ij, it includes walks that return to node i immediately after traversing edge ij, limiting their exploration of the network. We define such walks as **backtracking**, i.e. a walk is backtracking if it contains at least one node subsequence of the form iji. Otherwise, a walk is said to be **non-backtracking**.

Non-Backtracking Walk (NBTW) centrality addresses the issue of counting backtracking walks. It shares the foundational concept with Katz centrality but exclusively considers non-backtracking walks.

General formula. Denote by $P_k(A)$ a matrix whose (i,j)th entry is the number of nonbacktracking walks of length $k \geqslant 0$ from i to j. The NBTW centrality matrix is defined as

$$C_{\text{NBTW}}(A,t) = P_0(A) + tP_1(A) + t^2P_2(A) + t^3P_3(A) + \dots \qquad [2]$$

for suitable parameters t so that the above sum converges. For now, we assume $0 < t < 1$ to ensure that longer walks carry less weight, but the discussion on the radius of convergence will follow. The centrality measure for node i is then defined as the sum of the ith row of $C_{\text{NBTW}}(A,t)$:

$$\text{NBTW}_i(A,t) = \sum_{j=1}^{n} \left(C_{\text{Katz}}(A,t)\right)_{ij}.$$

Building upon the work in (7, 8), we aim to express the above formulas in a more computationally efficient manner. We will start by showing the following recurrence relation

$$P_k(A) = AP_{k-1}(A) - (D-I)P_{k-2}(A) \qquad [3]$$

for $k \geqslant 3$, where D is the diagonal matrix with (i,i)th entry equal to the degree of node i. The proof is by induction. For the base cases, $P_0(A) = I$ and $P_1(A) = A$ (as walks of length 1 are solely to the neighbors).

Additionally, $P_2(A) = A^2 - D$. Indeed, as shown in (18), A^2 counts walks of length 2 between nodes i and j. Since the only backtracking walks of length 2 are the ones of the form iri, where r is a neighbour of i, and there are $\deg(i)$ of them for each i, we account for them by subtracting D from A^2.

For $k \geqslant 3$, we observe that:

(i) $\left(AP_{k-1}(A)\right)_{ij} = \sum_{r=1}^{n} a_{ir}\left(AP_{k-1}(A)\right)_{rj}$ counts nonbacktracking walks of length k of the form $ir\dots j$, where r is some neighbor of i, and backtracking walks of the form $iri\dots j$, with backtracking occurring only at the first three nodes.

(ii) $\left((D-I)P_{k-2}(A)\right)_{ij} = \left(\deg(i)-1\right)\left(P_{k-2}(A)\right)_{ij}$ counts the number of backtracking walks of the form $iri\dots j$, with backtracking occurring only at the first three nodes.

Indeed, we fix a non-backtracking walk from i to j of length $k-2$ in $(P_{k-2}(A))_{ij}$ ways and then choose r from $\deg(i)-1$ neighbours of i that don't lie on the fixed non-backtracking walk, so that backtracking will only occur at the first three nodes.

Hence, subtracting $(D-I)P_{k-2}(A)$ from $AP_{k-1}(A)$ gives $P_k(A)$, thus proving equation 3. Notice that (omitting in the notation dependence on A and t):

$$t\,A\,C_{\mathrm{NBTW}} = tAP_0 + t^2AP_1 + t^3AP_2 + \dots$$
$$t^2\,(D-I)\,C_{\mathrm{NBTW}} = t^2(D-I)P_0 + t^3(D-I)P_1 + \dots$$

Using equation 3 and the definition of the NBTW centrality matrix 2, after some algebra, we obtain

$$M(t)\,C_{\mathrm{NBTW}}(A,t) = \left(1-t^2\right)I,$$

where $M(t) := I - t\,A + t^2\,(D-I)$ is called the **deformed graph Laplacian** associated with A. Therefore, if $\mathbf{v}_{\mathrm{NBTW}}$ is a centrality vector with the i-th entry equal to $\mathrm{NBTW}_i(A,t)$, it satisfies $M(t)\mathbf{v}_{\mathrm{NBTW}} = (1-t^2)\mathbf{1}$. This means that NBTW centrality may be computed at a similar cost as Katz centrality.

Radius of convergence. A detailed discussion of the radius of convergence for the power series in 2 can be found in Section 7 of (6). The key results are presented in Theorem 7.3 and Proposition 7.5 therein. While a thorough exploration of the radius of convergence is beyond the scope here, we offer a practical usage outline.

In network analysis, a natural range for t is $(0,1)$. However, as indicated in Section 8 of (6), given that the network has at least one connected component with average degree > 2 (which is often a case for real-world networks), $C_{\mathrm{NBTW}}(A,t)$ converges within the radius of convergence equal to $\mu = 1/\rho(C)$, where

$$C := \begin{pmatrix} A & I-D \\ I & 0 \end{pmatrix}$$

is the **companion linearization of rev**$M(t)$. Hence, in practice, we restrict t to the interval $(0, 1/\rho(C))$.

Limiting behaviour. Under the assumption of a connected network, according to Theorem 10.1. in (6), the rankings produced by NBTW centrality converge to:

(i) Degree centrality ranking when $t \to 0+$;

(ii) Nonbacktracking eigenvector centrality ranking (6, 23) when $\mu \leqslant 1$ and $t \to \mu-$.

2. Comparison of Katz and NBTW centrality measures

Theoretical comparison.

(i) NBTW and Katz centralities have similar computational complexities because they both require inverting an $n \times n$ matrix.

(ii) NBTW centrality has a larger radius of convergence than Katz centrality, as observed in (6).

(iii) NBTW and Katz centralities yield identical rankings when $\alpha, t \to 0+$. However, they give different rankings when their respective parameters approach their radii of convergence.

(iv) NBTW centrality, as highlighted in (6), may reduce localization (23), i.e. a disproportionate importance assignment to a limited set of nodes in the network. While localization is not always undesirable, this indicates a fundamental difference in the rankings implied by NBTW and Katz centrality measures. For a theoretical example of this phenomenon, refer to Section 11 in (6).

Comparison on real data. In this section, we analyse Katz and NBTW centrality measures using Twitter network data from five UK cities, sourced from (15). The dataset consists of five undirected networks representing Edinburgh, Glasgow, Cardiff, Bristol, and Nottingham, with node and edge counts detailed in Table 1. All networks are connected, and their average degrees are greater than 2. Therefore, we can apply previously discussed results on the radius of convergence and limiting behaviour of NBTW centrality.

Following (6, 12, 24), we use Kendall's τ correlation coefficient (25) to compare Katz and NBTW centrality measures. Furthermore, we investigate other comparison measures such as the inverse participation ratio to estimate the effects of localization (23) and overlap as proposed in (11). To further assess the localization effect, we introduce a new intuitive comparison measures, which is the percentage of nodes that contain 50% of the total mass.

Impact of the attenuation factor on rankings. Kendall's τ is a statistic measuring the degree of similarity between two data sets. Values close to 1 indicate considerable similarity between the rankings, whereas values near -1 suggest significant dissimilarity.

For each city, we compute Katz and NBTW centrality for 100 different values of $\alpha = t$, spanning from just above 0 to slightly below $1/\rho(A)$. For each fixed attenuation factor, we identify the top 5% of nodes based on their Katz centrality rankings. Subsequently, we find the respective rankings of these nodes according to NBTW centrality and determine Kendall's τ coefficient to quantify the similarity between these rankings.

We made two crucial decisions in our methodology. Firstly, we chose to compare Katz and NBTW centrality using the same attenuation factor, even though NBTW has a larger convergence radius than Katz. This decision has both advantages and disadvantages. On the one hand, we apply the same attenuation parameter to each walk length, which means they are penalized uniformly. However, because there are fewer nonbacktracking walks of a particular length compared to all walks of the same length, it makes intuitive sense to avoid harsh penalties for nonbacktracking walks. In this case, it might be sensible to opt for a larger factor. Despite these considerations, we chose to use the same attenuation factor. As shown in Table 1, $\rho(A)$ and $\rho(C)$ are very similar in magnitude for each city. Furthermore, using the same attenuation factor when comparing these two centrality measures seems reasonable for the sake of interpretation.

Secondly, we compare centrality measures for the top 5% of nodes rather than the entire network, following a similar approach to (6). In networks comprising approximately 2000 nodes, the top 5% may be considered the *big hitters* - the most crucial nodes whose rankings are of interest. To illustrate this heuristically, ranking an insignificant node at 1900 or 1950 place is less consequential for the interpretability than

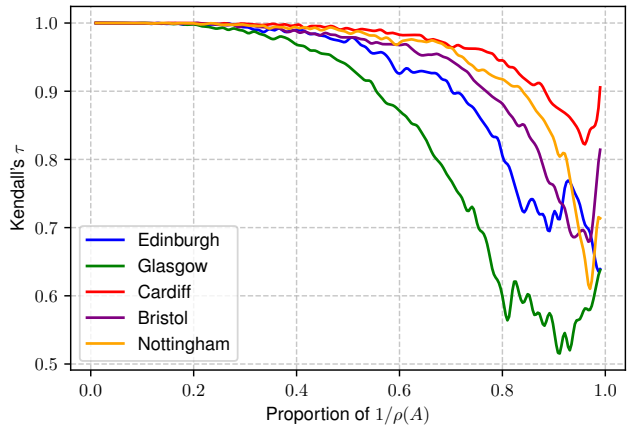

Fig. 1. Comparison of Katz(A, α) and NBTW(A, t) for the top 5% of ranked nodes with values of $\alpha = t$ ranging from slightly above 0 to just below $1/\rho(A)$.

classifying the most crucial node as 50th instead of 1st. Besides, computing Kendall's τ for all nodes may introduce noise from unimportant nodes, potentially impacting results.

The results, illustrated in Figure 1, reveal a consistent pattern across all cities. When α constitutes a small fraction of $1/\rho(A)$, Kendall's correlation coefficient between Katz and NBTW centralities is close to 1. It aligns with the theoretical expectation that both rankings converge to degree centrality rankings when the attenuation factor is close to 0.

As the attenuation factor increases, there is a decrease in Kendall's τ correlation coefficient, indicating disparate rankings between the two measures. There seems to be a negative correlation between the extent of Kendall's τ decrease and the magnitude of the spectral radius of the adjacency matrix, $\rho(A)$. Specifically, the decline in Kendall's τ for Cardiff, possessing the largest $\rho(A)$, is notably lower than for other cities with smaller spectral radii. The significant decrease in Kendall's τ for Glasgow is particularly noteworthy. A possible explanation for this can be found in Section 12 of (6).

Moreover, an intriguing phenomenon emerges as α approaches the radius of convergence, leading to a surprising increase in Kendall's τ. A heuristic explanation for this phenomenon is as follows. First, recall that the NBTW centrality measure satisfies the equation

$$M(t)\mathbf{v}_{\text{NBTW}} = (1 - t^2)\mathbf{1} \qquad [4]$$

where $M(t) = A - tA + t^2(D - I)$. When the spectral radius $\rho(A)$ is relatively large, the radius of convergence $1/\rho(A)$ becomes small. Consequently, by neglecting terms of order greater than 2, we simplify $M(t) \approx I - tA$ and $1 - t^2 \approx 1$. Thus, equation 4 can be rearranged to $(I - tA)\mathbf{v}_{\text{NBTW}} = \mathbf{1}$, which is also satisfied by the Katz centrality measure. Furthermore, the proximity of the radii of convergence for Katz and NBTW centrality measures across the five cities suggests a correlation between the limiting centralities of these centrality measures (i.e. eigenvector and nonbacktracking eigenvector centralities), particularly for the top 5% of nodes. However, it should be emphasized that as α approaches $1/\rho(A)$, the rankings are not identical but only seem to be positively correlated. Also, the argument above holds primarily when $\rho(A) \gg 1$ and should be solely treated as a heuristic explanation of the phenomenon in Figure 1.

Table 1. Comparison of Katz and NBTW centrality measures

	Edinb.	Glasg.	Card.	Brist.	Nott.
Nodes	1645	1802	2685	2892	2066
Edges	2146	2284	4444	4538	3155
$\rho(A)$	7.8	8.0	20.3	14.0	13.9
$\rho(C)$	5.5	5.1	18.8	12.1	12.3
IP-K	0.011	0.016	0.005	0.006	0.006
IP-NA	0.0022	0.0019	0.0027	0.0023	0.0025
IP-NC	0.01	0.0097	0.006	0.007	0.0062
IP-E	0.049	0.088	0.035	0.042	0.049
IP-NE	0.019	0.019	0.033	0.034	0.043
50-K	0.29	0.29	0.38	0.32	0.33
50-NA	0.37	0.38	0.41	0.37	0.38
KT-K_NA	0.734	0.606	0.914	0.839	0.891
KT-K_NC	0.815	0.643	0.958	0.864	0.885
O-K_NA	0.726	0.5	0.956	0.846	0.907
O-K_NC	0.822	0.765	0.971	0.882	0.943

IP: inv. part. ratio; 50: 50% measure; KT: Kendall's τ; O: overlap; K: Katz$(0.85/\rho(A))$; NA: NBTW$(0.85/\rho(A))$;NC: NBTW$(0.85/\rho(C))$; E: eigenvalue; NE: nonbacktracking eigenvalue

Localisation. While rankings provide valuable insights, they do not encompass the full picture when interpreting centrality results. Both measures assign certain weights to nodes, indicating their quantitative importance. As noted in (6), Katz centrality can exhibit localization, concentrating most weight on a few nodes. (23) proposes quantifying localization through an **inverse participation ratio** defined as

$$S = \sum_{i=1}^{n} v_i^4$$

for a vector $\mathbf{v} \in \mathbb{R}^2$ with $\|\mathbf{v}\|_2 = 1$ and $v_i \geqslant 0$. Higher values of the inverse participation ratio signify greater localization.

Table 1 shows that Katz centrality has a greater localization effect than NBTW centrality for the same attenuation factor. Evaluating the inverse participation ratio of NBTW centrality for $t = 0.85/\rho(C)$ suggests a localization effect comparable to Katz centrality for $\alpha = 0.85/\rho(A)$. However, this is due to using bigger attenuation factor for NBTW centrality in that case. Comparing the limiting rankings of the two centralities - eigenvalue and nonbacktracking eigenvalue centrality - reveals that the former exhibits more substantial localization. This aligns with findings in (23).

Another argument supporting the view that Katz centrality demonstrates a stronger localization effect than the NBTW centrality comes from the proportion of nodes that contain 50% of the total weight. In both cases, the proportion is lower for the Katz centrality measure, indicating that the weight is more concentrated for that measure.

Overlap. We can also examine the overlap (11), defined as $|K \cap N|/|K \cup N|$, where K is the set of the top 5% nodes from Katz $(0.85/\rho(A))$, and N is the set of the top 5% nodes from NBTW $(0.85/\rho(A))$ or NBTW $(0.85/\rho(C))$. As observed in Table 1, the overlap between these two measures is approximately 80% (with a notable exception in Glasgow), slightly higher when comparing Katz $(0.85/\rho(A))$ and NBTW $(0.85/\rho(C))$.

In summary, examining the overlap between Katz and NBTW reveals agreement on roughly 80% of the top 5% nodes, with overlap increasing for a larger value of the NBTW attenuation factor. Kendall's τ further confirms that the rankings of these two centralities are positively correlated.

The main distinction between these two measures is that, unlike Katz centrality, NBTW centrality does not allocate such a substantial amount of weight to a small number of nodes but appears to distribute it more evenly among the most significant nodes.

3. Network Errors

Edge errors. As previously mentioned, network centrality measures can contain errors. In the examples above, we used reciprocated Twitter mentions networks for five UK cities, and the data collection method is explained in Section 4 of (15). Investigating a social network over a long period is beneficial. Since the data covers only 28 days, extending this duration could uncover more reciprocated tweets. It implies that some edges could be missing from the network, leading to potential false-negative edge errors in the data. Additionally, the Twitter network is susceptible to false-positive edges and nodes. According to (26), between 9% and 15% of active Twitter accounts may be bots. In this section, we will examine how the incidence of false-negative and false-positive edges affects Katz and NBTW centrality measures.

Methodology for estimating the robustness of centrality measures. Our approach to assessing the robustness of Katz and NBTW centrality measures in the presence of false-positive and false-negative edges follows the methods outlined in (11–13). Let us briefly summarise the methodology for estimating robustness from (12).

We use M to denote the measured network and H to denote the hidden network, representing the true (but unknown!) network. We calculate the robustness of centrality measure c by $\rho_c(M, H)$, where ρ is some robustness measure. We will use Kendall's τ between the rankings of centrality measures for M and H and a 5% overlap, as defined in the previous section. There are alternative ways to measure the robustness of a system, which have been suggested in (11). For instance, we can consider the proportion of times the top node of M is in the top 1, top 3, or top 10% of nodes in H.

As H is unknown, we cannot determine the exact value of the true robustness $\rho_c(M, H)$. However, we can estimate it using the method suggested in (12):

$$\widehat{\rho}_c(M, H) = \mathbb{E}\left[\rho_c\left(M, \varphi(M)\right)\right]. \qquad [5]$$

$\varphi(M)$ denotes a random graph created by introducing errors to M, such as adding randomly new edges (false-positive) or randomly deleting some edges (false-negative).

In practice, we estimate the expectation in equation 5 using the Monte Carlo method. We independently apply the error mechanism multiple times to the measured network and report the average impact of this procedure as an estimate of the robustness. In our numerical computations, we simulated 10 *'paths'* for each type of error (false-positive and false-negative edges). For each path, we calculated Kendall's τ correlation coefficient and the 5% overlap between the measured and the erroneous network and took the mean.

Numerical results. First, we use Kendall's tau correlation coefficient as a measure of robustness, with an attenuation factor of $0.8/\rho(A)$ for Katz centrality and $0.8/\rho(C)$ for NBTW centrality (we update $\rho(A)$ and $\rho(C)$ each time we delete or add edges). Figure 2 depicts the results for Edinburgh, with

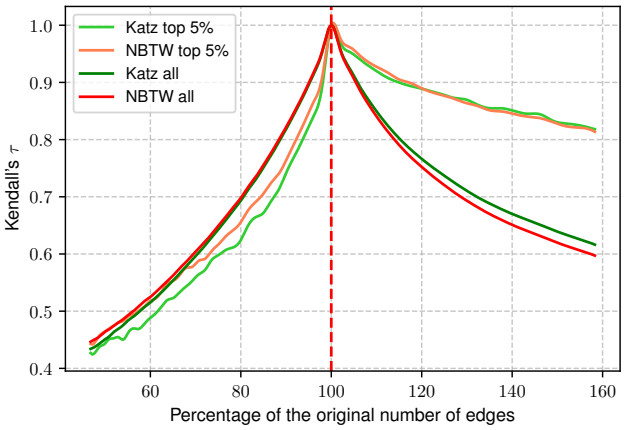

Fig. 2. Behaviour of Kendall's τ in response to network error for Edinburgh.

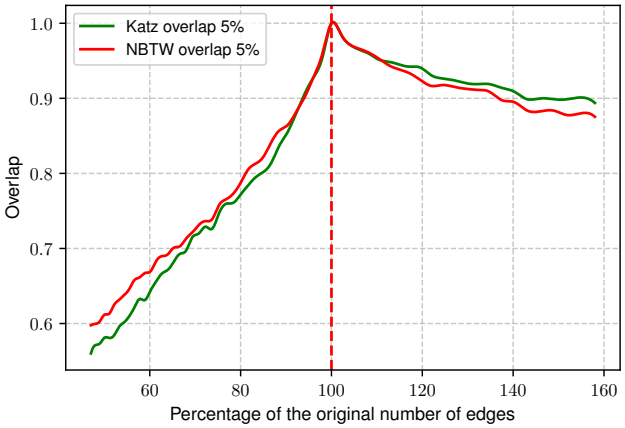

Fig. 3. Behaviour of 5% overlap in response to network errors for Bristol.

Kendall's τ correlation coefficient calculated for the top 5% of ranked nodes and the entire network. The behaviour of the correlation coefficient for other cities is similar to that of Edinburgh.

We find that both Katz and NBTW centrality measures exhibit similar robustness to false-negative and false-positive edges. Indeed, the relevant graphs follow a virtually identical pattern. It is similar to Borgatti's (11) observations that the four therein considered centrality measures (degree, closeness, betweenness, and eigenvector centralities) "are surprisingly similar in terms of pattern and level of robustness".

We find that false-positive errors have less of an impact on the rankings of the top 5% of nodes than false-negative edges. Even if we add 60% of randomly generated new edges, the ranking is not significantly distorted, and Kendall's τ correlation coefficient remains at around 0.85 for all networks. However, when calculating Kendall's τ for all nodes in the network, the false-positive error has a more significant effect on the ranking of the whole network. This is because the results are more susceptible to noise from nodes that are ranked lower. Therefore, it is safe to say that even with false-positive errors, the top-ranked nodes are still well-classified and well-ranked.

It seems that Kendall's τ decreases linearly as edges are removed for both centrality measures and both examined

numbers of nodes. This trend could be because the analysed networks are sparse, with densities around 0.001. In sparse networks, removing even a small percentage of edges, such as $5\% - 10\%$, might lead to a catastrophic impact on walk-based centralities, as the network becomes disconnected rather quickly. The impact of network density on the robustness of centrality measures is further discussed in (11).

Recall that the 5% overlap is defined as $|M \cap \varphi(M)|/|M \cup \varphi(M)|$. Figure 3 illustrates the impact of false-positive and false-negative errors on the overlap in the Bristol network. The same patterns were observed in all cities. Additionally, the reduction in overlap follows a similar pattern and level of degradation as Kendall's correlation coefficient. The overlap gradually decreases with false-positive edges but rapidly declines with false negatives. This observation further confirms that the rankings of the most important nodes from Katz and NBTW centrality measures are relatively robust in the presence of false-positive edges, at least for sparse networks.

Discussion of limitations and possible future avenues

While NBTW centrality measure addresses the issue of counting redundant walks by Katz centrality that immediately return to the previously visited node, it still considers walks that endlessly cycle through a short cycle of length 3 or 4. (27) deals with this issue by introducing non-cyclic centrality measures. A comparison of these measures with NBTW centrality would be valuable.

Furthermore, the results of both Katz and NBTW centrality measures rely on a choice of attenuation factor. According to (21), a reasonable choice for the attenuation factor is 0.85 of the radius of convergence. It is supported by an example from (6), where the higher attenuation factors lead to better performance in including essential proteins among the top-ranked nodes. There is no definitive answer to the optimal choice of parameter, but the results presented in (28) provide valuable insights for making intelligent choices for the attenuation factor. (28) also argues that, in most cases, the rankings are relatively stable and do not change significantly for different choices of the attenuation factor, even if the actual scores vary widely.

Another interesting question to consider is how to measure the effect of localisation. The inverse participation ratio is a well-studied measure (23), and there are heuristic results about its assessment of localisation when the number of nodes increases. It would be interesting to see whether a measure that considers the allocation of 50% of the mass would yield similar results to those found for the inverse participation ratio. A possible improvement of this measure is to consider another percentage than 50% or an adaptable benchmark that doesn't require fixing a certain proportion of mass. However, this requires further research.

We believe that there is room for improvement when it comes to assessing the impact of network errors on centrality measures. While Kendall's τ and overlap provide an intuitive indication of the degradation of analysed centrality measures, they may not be effective in the presence of false-negative and false-positive nodes. Therefore, it would be beneficial to have another method of assessing robustness. Performing non-parametric statistical tests and reporting p-values could indicate at what point the introduction of errors results in a statistically significant difference in rankings.

Materials and Methods

The report draws ideas from multiple research articles as referenced throughout the text. Figures and numerical results were generated using Jupyter Notebooks and Python 3.9.6, with key packages including NetworkX for network implementation, Matplotlib for graph production, and Numpy for matrix manipulations.

A significant computational challenge was recalculating the spectral radii $\rho(A)$ and $\rho(C)$ when introducing random errors to the network. We addressed this challenge in the following ways:

(i) (16) shows that $\rho(C)$ remains unchanged when we remove leaves from the network (this isn't true for $\rho(A)$). This operation, known as pruning, can be performed recursively, reducing computation for a typical network size by around 30%, as noted in (6).

(ii) We implemented the power iteration method (29) to efficiently find the spectral radii of matrices. Given the sparsity of the networks, this algorithm converges reasonably quickly.

We are grateful to the authors of (12) for sharing their code online. Their ideas helped implement false-positive and false-negative error mechanisms. The code used in this project is available upon request.

ACKNOWLEDGMENTS. I would like to thank Prof. Peter Grindrod CBE for providing me with the knowledge without which this report would not have been possible.

1. C Laghridat, M Essalih, A set of measures of centrality by level for social network analysis. *Procedia Comput. Sci.* **219**, 751–758 (2023).
2. A Landherr, B Friedl, J Heidemann, A critical review of centrality measures in social networks. *Wirtschaftsinformatik* **52**, 367–382 (2010).
3. D Koschützki, F Schreiber, Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio* **2**, 193–201 (2008).
4. A Soofi, et al., Centrality analysis of protein-protein interaction networks and molecular docking prioritize potential drug-targets in type 1 diabetes. *Iran J Pharm Res* **19**, 121–134 (2020).
5. L Katz, A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
6. P Grindrod, DJ Higham, V Noferini, The deformed graph laplacian and its applications to network centrality analysis. *SIAM J. on Matrix Analysis Appl.* **39**, 310–341 (2018).
7. H Stark, A Terras, Zeta functions of finite graphs and coverings. *Adv. Math.* **121**, 124–165 (1996).
8. A Tarfulea, R Perlis, An ihara formula for partially directed graphs. *Linear Algebr. its Appl.* **431**, 73–85 (2009).
9. D Brewer, C Webster, Forgetting of friends and its effects on measuring friendship networks. *Soc. Networks* **21**, 361–373 (2000).
10. D Wang, X Shi, D Mcfarland, J Leskovec, Measurement error in network data: A re-classification. *Soc. Networks* **34** (2012).
11. S Borgatti, K Carley, D Krackhardt, On the robustness of centrality measures under conditions of imperfect data. *Soc. Networks* **28**, 124–136 (2006).
12. C Martin, P Niemeyer, Influence of measurement errors on networks: Estimating the robustness of centrality measures. *Netw. Sci.* **7**, 180–195 (2019).
13. ZW Almquist, Random errors in egocentric networks. *Soc. Networks* **34**, 493–505 (2012).
14. E Costenbader, T Valente, The stability of centrality measures when networks are sampled. *Soc. Networks* **25**, 283–307 (2003).
15. P Grindrod, TE Lee, Comparison of social structures within cities of very different sizes. *Royal Soc. Open Sci.* **3**, 150526 (2016).
16. F Arrigo, P Grindrod, D Higham, V Noferini, Non-backtracking walk centrality for directed networks. *J. Complex Networks* **6**, 54–78 (2018).
17. F Arrigo, DJ Higham, V Noferini, R Wood, Weighted enumeration of nonbacktracking walks on weighted graphs (2023).
18. A Duncan, Powers of the adjacency matrix and the walk matrix. *The Collect.* **9**, 4–11 (2004).
19. CD Meyer, *Matrix Analysis and Applied Linear Algebra.* (Society for Industrial and Applied Mathematics, USA), (2000).
20. R Plemmons, M-matrix characterizations.i—nonsingular m-matrices. *Linear Algebr. its Appl.* **18**, 175–188 (1977).
21. M Benzi, C Klymko, On the limiting behavior of parameter-dependent network centrality measures. *SIAM J. on Matrix Analysis Appl.* **36**, 686–706 (2015).
22. P Bonacich, Factoring and weighting approaches to status scores and clique identification. *The J. Math. Sociol.* **2**, 113–120 (1972).
23. T Martin, X Zhang, MEJ Newman, Localization and centrality in networks. *Phys. Rev. E* **90**, 052808 (2014).
24. D Schoch, T Valente, U Brandes, Correlations among centrality indices and a class of uniquely ranked graphs. *Soc. Networks* **50**, 46–54 (2017).
25. MG Kendall, A new measure of rank correlation. *Biometrika* **30**, 81–89 (1938).
26. O Varol, E Ferrara, C Davis, F Menczer, A Flammini, Online human-bot interactions: Detection, estimation, and characterization. *Proc. Int. AAAI Conf. on Web Soc. Media* **11** (2017).
27. F Arrigo, DJ Higham, V Noferini, Beyond non-backtracking: non-cycling network centrality measures. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **476**, 20190653 (2020).
28. CF Klymko, Ph.D. thesis (Emory University, Atlanta, GA) (2013).
29. BN Parlett, *The symmetric eigenvalue problem.* (SIAM), (1998).