Analyzing the Impact of Data Accuracy on Katz and NBTW Centrality Measures in Social Media Network

1081913

This manuscript was compiled on January 7, 2024

This study embarks on an in-depth analysis of centrality measures in network data, with a particular focus on Twitter reciprocated mentions networks across UK cities, including Edinburgh, Glasgow, Cardiff, Bristol, and Nottingham. Our primary objective is to compare Katz centrality and Non-Backtracking Walk (NBTW) centrality measures. Then, we assess their sensitivity to data inaccuracies, specifically false positives and false negatives.

Employing datasets representing Twitter interactions from the five cities, we calculate both the Katz and NBTW centrality matrices. To address the varying edge densities, we normalize these matrices before analysis. Sensitivity to data errors was investigated by introducing controlled false positives and negatives, and the impact on centrality measures was quantified using Mean Squared Error (MSE).

Our analysis revealed significant differences in how Katz and NBTW centrality measures respond to changes in network structure, particularly under the influence of false data. Notably, false negatives, which mean missing edges, displayed a more substantial impact on centrality measures than false positives (successive edges), even after normalization. Additionally, the MSE between Katz and NBTW centralities varied notably across different cities. The pronounced effect of false negatives on centrality distributions, in particular, calls for careful consideration of data completeness in network studies.

Significance Statement

This report critically contrast the Katz centrality matrix and NBTW centrality matrix, and examines the sensitivity of them to data accuracy, specifically addressing and comparing the impact of false positives and negatives in network analysis. Utilizing Twitter data from UK cities, the report reveals significant differences in how these centrality measures respond to two types of data errors, even after normalization. This work contributes to a deeper understanding of the twocentrality measures in complex networks, especially the similarities and differences between them.

Network Analysis|Computational Network Science | Katz Centrality | Non-Backtracking Walk (NBTW) Centrality | Sensitivity

In the contemporary era of digital communication and social media, network analysis has become an invaluable tool for understanding complex systems. It has been applied in both artificial and biological social networks as well as organisms(1)(2)(3), the study of networks offers insights into the structural and dynamic properties of various interconnected systems.(4) Central to this field is the analysis of nodes (or vertices) and edges that form the crux of these networks.(5) Centrality measures in network analysis are pivotal in identifying the most influential nodes within a network, aimed at identifying the most influential or important nodes within a network.(6) While these measures often provide insights into the a quantitative assessment of nodes' relative importance, interpreting the raw centrality matrices can be challenging, particularly when attempting to compare different centrality measures across a network.(7)

Katz centrality, introduced by Leo Katz in 1953(8), is a measure that extends the concept of degree centrality by considering not just the immediate neighbors of a node but also nodes that are further away but connected. It is calculated based on the number of all possible walks between node pairs, inversely weighted by their length, thus incorporating a global view of the network.

In contrast to Katz centrality, NBTW centrality provides a unique perspective by focusing on non-backtracking walks – paths where a walk does not immediately reverse its direction. (9)This centrality measure offers a more nuanced view of node influence by eliminating redundancy in path counting. (7)

Though network analysis has been increasingly developed, there is relatively little discussion about the false positives (erroneous edges) and false negatives (missing edges), which are common data inaccuracies that can significantly distort network analysis outcomes. (10) Understanding the sensitivity of centrality measures to these errors is crucial for ensuring the reliability of network studies. (5) Thus, this report aims to explore the comparative analysis of Katz centrality and NBTW centrality, particularly focusing on their sensitivity to data errors. By examining the impact of false positives and false negatives on these centrality measures, we seek to

Please provide details of author contributions here.

Please declare any competing interests here.

[1]A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

provide valuable insights into the robustness and reliability of network analysis in the face of data inaccuracies.

1. Centrality Matrices

Consider a large sparse undirected network on n (large) vertices, with symmetric adjacency matrix A, which has zeros along its diagonal, which means that there are no self-loops. Denote the matrix that counts the number of walks of length r between all pairs of vertices as P_r, that is, $(P_r)_{ij}$ is exactly the number of walks of length r between vertex i and vertex j. Then we have and $P_r = AP_{r-1}, r \geq 1$, and $P_0 = I$. Then the Katz Centrality Matrices is defined as the Generating Function $Q = \sum_{r=0}^{\infty} \alpha^r P_r$, where α should be small enough so that Q can converges.

Since the adjacency matrix A of our undirected graphs is real and symmetric thus diagonalizable, it can be represented as PDP^{-1}, so

$$\begin{aligned} Q &= \sum_{r=0}^{\infty} \alpha^r P_r \\ &= I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \cdots \\ &= P(I + \alpha D + \alpha^2 D^2 + \alpha^3 D^3 + \cdots)P^{-1} \\ &= P(I - \alpha D)^{-1} P^{-1} \\ &= (PIP^{-1} - P(\alpha D)P^{-1})^{-1} \\ &= (I - \alpha A)^{-1} \end{aligned} \quad [1]$$

, where α is required to be small enough to ensure that Q is well-defined, which means that $\alpha < \rho(A)$, the spectral radius of A

The generating function for Non-Backtracking Walk, Q, which is called NBTW centrality matrix, is given by $\sum_{r=0}^{\infty} \alpha^r P_r$ where P_r is the number of non-backtracking walks of length r. For $r \geq 3$, we have

$$P_r = AP_{r-1} - (D - I)P_{r-2}$$

, given that AP_{r-1} extends each non-backtracking walk of length $r-1$ by one step along an edge in the network to length r and that $(D-I)P_{r-2}$ counts walks of length $r-2$ that are extended by two steps, where D is the diagonal matrix of vertex degrees and I is the identity matrix, the first step is to any neighbor and the second step is back to the original vertex, thereby forming a backtracking walk which is not allowed and should be removed.

since Q is a matrix valued function, for some real α such that Q is well defined,

$$\begin{aligned} Q &= P_0 + \alpha P_1 + \alpha^2 P_2 + A\sum_{r=3}^{\infty} \alpha^r P_{r-1} - (D-I)\sum_{r=3}^{\infty} \alpha^r P_{n-2} \\ &= I + \alpha A + \alpha^2(A^2 - D) + \alpha A(Q - I - \alpha A) \\ &\quad - \alpha^2(D-I)(Q-I) \\ &= (I + \alpha A + \alpha^2 A^2 - \alpha^2 D - \alpha A - \alpha^2 A^2 + \alpha D^2 - \alpha^2 I) \\ &\quad + (\alpha A - \alpha^2(D-I))Q \\ &= (1-\alpha^2)I + (\alpha A - \alpha^2(D-I))Q. \\ \therefore &(I - \alpha A + \alpha^2(D-I))Q = (1-\alpha^2). \end{aligned}$$

So we have

$$Q = (1-\alpha^2)M(\alpha)^{-1} \quad [2]$$

where

$$M(\alpha) = (I - \alpha A + \alpha^2(D-I))$$

and α must chosen to ensure that $M(\alpha)$ is non-singular. Fortunately, both $I - \alpha A$ and $\alpha^2(D-I)$ are positive definite if $0 \geq \alpha < \rho(A)^{-1}$, thus $M(\alpha)$ is positive definite, which meets the requirement. .

2. Contrasting two centrality matrices

In this section, we tried to contrast the Katz centrality matrix and the NBTW centrality matrix on our sample Twitter reciprocated mentions networks. We firstly load the Twitter reciprocated mentions networks of five UK cities, i.e. Edinburgh, Glasgow, Cardiff, Bristol and Nottingham, as our samples. The visualization of the fives networks is shown in the middle row of the figure 1.

Then, we compute the two central matrices given by 1 and 2 for each city in turn and for α ranging from very close to 0 to very narrow to $\rho(A)$, and try to visualize them as graphs, which is shown in the figure 2 (Upper row are Katz centrality matrices while Lower row are NBTW centrality matrices).

As the figure 2 shows, it is very hard to differ these two measures by the matrix graph. To address this challenge, we employed a method of summing the rows of both the Katz centrality matrix and the NBTW (Non-Backtracking Walk) centrality matrix. This approach transforms the complex centrality matrices into more accessible centrality distributions, simplifying the comparative analysis.

The primary justification for this approach lies in its ability to distill the essence of each centrality measure into a form that is both analytically and visually comprehensible. By summing the rows of each matrix, that is,

$$\begin{aligned} C_{Katz} &= Q_{Katz}\mathbf{1} \\ C_{NBTW} &= Q_{NBTW}\mathbf{1} \end{aligned}$$

where C_{Katz} and C_{NBTW} are centrality distribution of networks, respectively. They that reflects the overall centrality landscape of the network.

We test the MSE between the Katz centrality distribution and the NBTW centrality distribution.The result is presented in the table 1. Interestingly, within the context of our sample data from the Twitter networks of five UK cities, there were no substantial differences between the Katz and NBTW centrality distributions. This observation was somewhat unexpected, given the theoretical distinctions between these measures.

This result can imply that, for the specific structure and dynamics of samples, both centrality measures converge in identifying influential nodes, despite their methodological differences, especially when applied to large and sparse networks like our samples.

3. Sensitivity Analysis of Centrality Measures to Data Errors

In network analysis, the accuracy of data is paramount, as errors like false positives and false negatives may significantly alter the interpretation of a network's structure and dynamics. In networks, false positives are edges which are in the network which really should not have been there and false negatives are edges missed. Given the practical inevitability of such errors

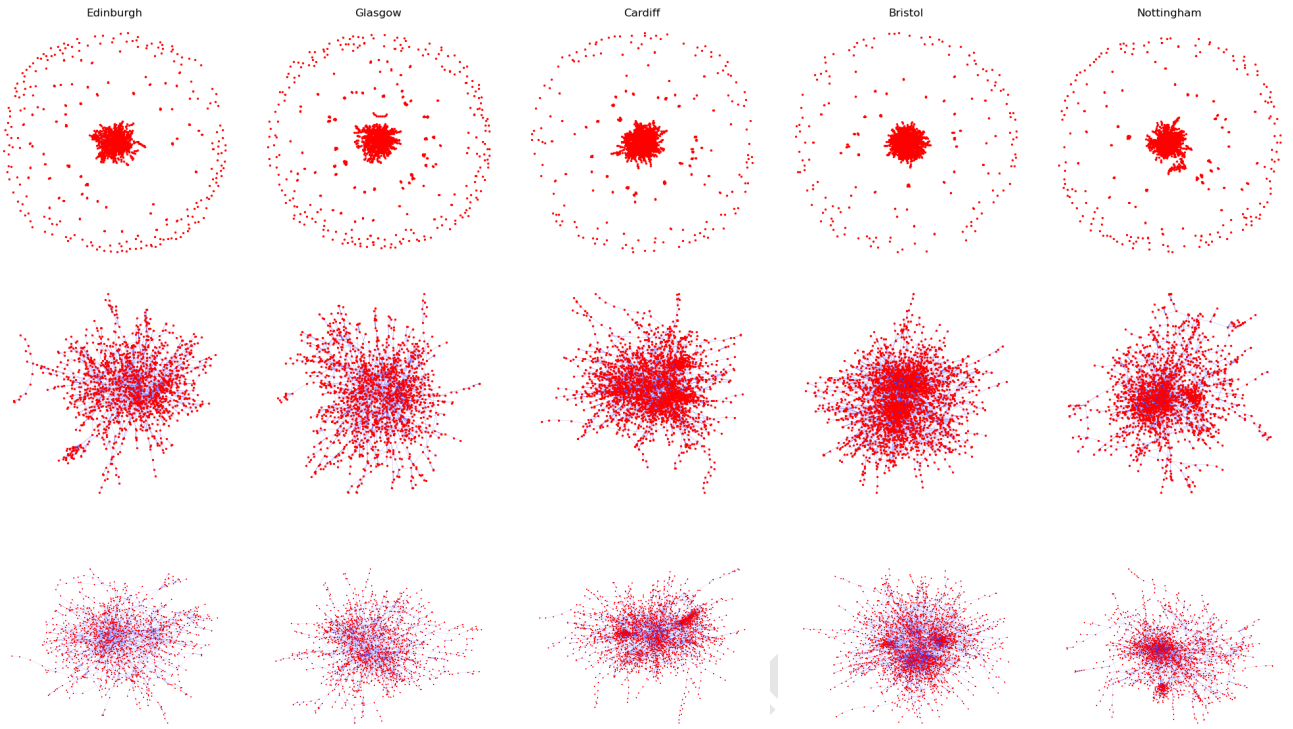

Fig. 1. The first row shows the networks inducing false negative errors, while the second row shows the original networks and the third row shows the networks inducing false positive errors..

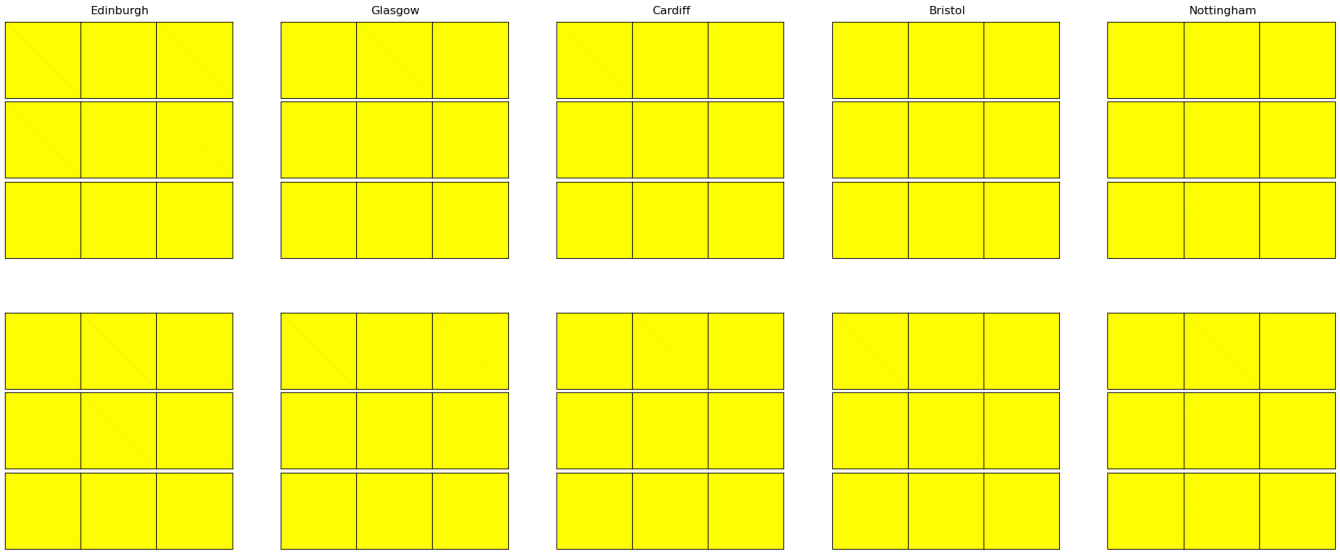

Fig. 2. Upper row are Katz centrality matrices while Lower row are NBTW centrality matrices, for α ranging from $0.1\rho(A)$ to $0.9\rho(A)$

in real-world data, understanding their impact on centrality metrics is crucial for network sensitivity analysis.

Our approach involvs introducing controlled levels of false positives (adding edges) and false negatives (deleting edges) separately into the Twitter reciprocated mentions networks of Edinburgh and Glasgow, which are initially considered as 'ground truth'. This process simulated the common inaccuracies encountered in network data collection, ranging from mild to severe levels of error. Each altered network was then analyzed to compute its Katz centrality and NBTW centrality matrices.

A. Normalization Approach. Due to the alterations in edge density resulting from the induced errors, as visualized in the figure 1 (the first row shows the networks inducing false negative errors, while the second row shows the original networks and the third row shows the networks inducing false positive errors). We can obviously see that inducing

Table 1. The MSE between the Katz centrality and NBTW centrality in the five cities

MSE	$\alpha\rho(A)$	Edin	Glas	Card	Bris	Nott
$(\times 10^{-11})$	0.1	6.018	6.082	0.2731	1.160	1.226
$(\times 10^{-10})$	0.2	15.60	16.83	0.6553	3.115	3.294
$(\times 10^{-9})$	0.3	13.62	15.50	0.5276	2.780	2.895
$(\times 10^{-8})$	0.4	8.085	9.582	0.2871	1.649	1.667
$(\times 10^{-8})$	0.5	41.73	50.98	1.349	8.253	7.970
$(\times 10^{-7})$	0.6	21.75	27.12	0.6361	4.011	3.618
$(\times 10^{-6})$	0.7	13.31	16.77	0.3500	2.181	
$(\times 10^{-5})$	0.8	12.23	15.44	0.2894	1.704	1.157
$(\times 10^{-r})$	0.9	33.76	42.37	0.7338	4.012	2.098

errors significantly change the edge density as well as the whole structure of networks, which may affect the central measures. Thus, it was necessary to normalize the centrality distributions for a fair comparison. This normalization was achieved by dividing the centrality values by the respective edge density of the error-induced network. This step ensured that the centrality values were adjusted for the size and density changes in the network, allowing for a more accurate comparison with the original network's centrality distributions. By adjusting the centrality values to the edge density of each network, we achieved a more equitable basis for comparing centrality distributions across varying network conditions.

B. Mean Squared Error (MSE) Calculation. To quantitatively assess the impact of the induced errors, we calculated the Mean Squared Error (MSE) between the normalized centrality distributions of the error-induced networks and the original, error-free networks. The MSE provided a clear metric to evaluate the sensitivity of each centrality measure to the presence of false positives and false negatives.

From the figure 3, where the first row shows results for Edinburgh and the other row shows results for Glasgow. We can numerically analize the impact of false positives and false negatives.

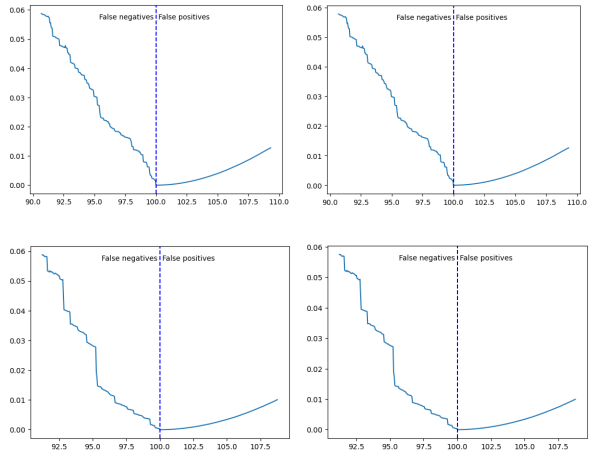

Fig. 3. MSE induced by errors for Edinburgh (first row) and Glasgow (secend row), where y-axis is the MSE and x-axis is the percentage of the edge density compared to the original networks

4. Results

Our analysis commenced with a comparative assessment of the Katz and Non-Backtracking Walk (NBTW) centrality measures across the Twitter networks of Edinburgh and Glasgow. The results are shown in the figure 3, where y-axis is the MSE and x-axis is the edge density compared to the original networks.

As shown in the figure, the MSE was consistently higher in networks with false negatives across both sample cities, which illustrates the centrality distributions were more significantly impacted by false negatives compared to false positives. The introduction of false positive edges (adding edges) generally led to an relatively small inflation in centrality scores across the network. Katz centrality exhibited a slightly higher sensitivity to these errors. Conversely, the removal of true edges (false negatives) generally resulted in a very significant effect of the centrality score, while the NBTW centrality shows a marginal resilience to this type of error, likely owing to its focus on non-backtracking paths..

Our results provide compelling evidence of the differential impacts of data inaccuracies on centrality measures in network analysis. The greater sensitivity of both Katz and NBTW centrality to false negatives, particularly in certain cities, displaying the importance of data completeness and accuracy in network studies. These findings not only help to understanding of centrality measures but also exhibit the necessity for rigorous data preprocessing and validation in network analysis.

5. Discussion

Our comparative analysis of Katz and NBTW centrality measures across different Twitter networks revealed key differences in how these measures interpret node influence. The distinct methodologies underlying Katz and NBTW centrality led to variations in node ranking, particularly pronounced at higher values of the scaling factor $\alpha\rho(A)$. This suggests that while Katz centrality may be more sensitive to short-path structures in the network, NBTW centrality provides a nuanced understanding of influence, excluding redundant pathways.

The study's findings regarding the differential impact of false positives and false negatives are particularly enlightening. The greater sensitivity of centrality measures to false negatives points to the critical role of network connectivity in determining node influence. This aligns with the understanding that the removal of edges can significantly disrupt network dynamics, more so than the addition of spurious connections.

Normalization proved to be an effective method for mitigating disparities caused by variations in network size and edge density. However, its limited ability to fully compensate for the distortions caused by false negatives highlights a crucial aspect of network analysis – the importance of data completeness and accuracy. This emphasizes the need for rigorous data validation and error correction in preprocessing stages of network analysis.

The city-specific trends observed in the study emphasize the importance of context in network analysis. Differences in network structures and interaction patterns across the cities suggest that centrality measures might not be universally applicable without consideration of the specific network

characteristics, for instance, matrices of our samples are all very large and sparse. This raises questions about the generalizability of centrality measures across different types of networks.

This research opens several avenues for future exploration. One key area is the development of centrality measures or analytical techniques that are more resilient to data inaccuracies, especially in the era of big data where such inaccuracies are commonplace. Another area of interest could be the exploration of centrality measures in networks with different topologies, such as weighted or directed networks, to understand the broader applicability of our findings.

The findings highlight the importance of considering the specific properties and limitations of centrality measures when interpreting their results in network analysis. This research contributes to a more nuanced understanding of network dynamics and offers a foundation for more accurate and robust network analysis methodologies.

Materials and Methods

Materials and Methods

Data Collection: The data for this study was sourced from Twitter reciprocated mentions networks of five UK cities: Edinburgh, Glasgow, Cardiff, Bristol, and Nottingham. The datasets represent interactions (mentions) between Twitter accounts within each city during a specified period.

Network Representation: Each city's Twitter interactions were represented as an undirected graph, where nodes correspond to Twitter accounts and edges represent mutual mentions. The adjacency matrix A of each network was constructed, ensuring no self-loops, to facilitate centrality analysis.

Centrality Measures Computation, including:

1. Katz Centrality: Calculated for each network using the formula $Q = (I - \alpha A)^{-1}$, where α is a scaling factor less than the spectral radius of A, ensuring convergence.

2. Non-Backtracking Walk (NBTW) Centrality: Determined using the deformed graph Laplacian method, with the centrality matrix given by $Q = (1 - \alpha^2)M(\alpha)^{-1}$, where $M(\alpha) = I - \alpha A + \alpha^2(D - I)$ and D is the diagonal matrix of node degrees.

Error Induction and Normalization: Controlled errors were introduced into the network data to simulate false positives (additional edges) and false negatives (missing edges). Subsequently, centrality distributions were normalized against the edge density of each altered network to account for changes in network size and connectivity.

Quantitative Analysis: The sensitivity of centrality measures to the induced errors was quantitatively evaluated using Mean Squared Error (MSE). MSE was calculated between the centrality distributions of the original and error-induced networks for each city.

Statistical Methods

Descriptive statistics were employed to analyze centrality distributions and MSE results. Comparative analysis was conducted to examine the variations in centrality measures across different cities and under varying conditions of error induction.

1. A Dongare, R Kharde, AD Kachare, , et al., Introduction to artificial neural network. *Int. J. Eng. Innov. Technol. (IJEIT)* **2**, 189–194 (2012).
2. E Skrimizea, Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies (2021).
3. G Muzio, L O'Bray, K Borgwardt, Biological network analysis with deep learning. *Briefings bioinformatics* **22**, 1515–1530 (2021).
4. BS Anderson, C Butts, K Carley, The interaction of size and density with graph-level indices. *Soc. networks* **21**, 239–267 (1999).
5. ZW Almquist, Random errors in egocentric networks. *Soc. networks* **34**, 493–505 (2012).
6. S Oldham, et al., Consistency and differences between centrality measures across distinct classes of networks. *PloS one* **14**, e0220061 (2019).
7. 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).
8. L Katz, A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
9. L Torres, KS Chan, H Tong, T Eliassi-Rad, Nonbacktracking eigenvalues under node removal: X-centrality and targeted immunization. *SIAM J. on Math. Data Sci.* **3**, 656–675 (2021).
10. SP Borgatti, KM Carley, D Krackhardt, On the robustness of centrality measures under conditions of imperfect data. *Soc. networks* **28**, 124–136 (2006).