Applied Network Science

RESEARCH Open Access

Map equation centrality: community-aware centrality based on the map equation

Christopher Blöcker[1*], Juan Carlos Nieves[2] and Martin Rosvall[1]

*Correspondence: christopher.blocker@umu.se

[1] Integrated Science Lab, Department of Physics, Umeå University, 901 87 Umeå, Sweden
[2] Department of Computing Science, Umeå University, 901 87 Umeå, Sweden

Abstract

To measure node importance, network scientists employ centrality scores that typically take a microscopic or macroscopic perspective, relying on node features or global network structure. However, traditional centrality measures such as degree centrality, betweenness centrality, or PageRank neglect the community structure found in real-world networks. To study node importance based on network flows from a mesoscopic perspective, we analytically derive a community-aware information-theoretic centrality score based on network flow and the coding principles behind the map equation: map equation centrality. Map equation centrality measures how much further we can compress the network's modular description by not coding for random walker transitions to the respective node, using an adapted coding scheme and determining node importance from a network flow-based point of view. The information-theoretic centrality measure can be determined from a node's local network context alone because changes to the coding scheme only affect other nodes in the same module. Map equation centrality is agnostic to the chosen network flow model and allows researchers to select the model that best reflects the dynamics of the process under study. Applied to synthetic networks, we highlight how our approach enables a more fine-grained differentiation between nodes than node-local or network-global measures. Predicting influential nodes for two different dynamical processes on real-world networks with traditional and other community-aware centrality measures, we find that activating nodes based on map equation centrality scores tends to create the largest cascades in a linear threshold model.

Keywords: Community-aware, Centrality, Map equation, Random walk, Huffman coding

Introduction

Networks are simple yet powerful representations of how things connect: the world wide web captures connections between websites, and social networks describe relationships between persons. So-called centrality measures determine node importance and enable us to rank nodes, compare them with each other, and find the most important ones. Real-world applications are manifold and include identifying the most popular websites, which components in an infrastructure network have the most impact when they fail, and who drives disease spreading in a social network.

Classical centrality measures consider node importance on a microscopic scale at the node level or on a macroscopic scale at the network level. For example, degree centrality defines a node's importance proportional to its degree, and betweenness centrality calculates node importance as the number of shortest paths that pass through it (Koschützki et al. 2005). Eigenvector centrality-based measures, such as Katz centrality (Katz 1953) and PageRank (Gleich 2015), implement a reputation system and derive a node's importance from how important its neighbours are, leading to a system of recursive equations. However, real-world networks often exhibit community structure. Loosely speaking, they contain groups of nodes, so-called communities, with more connections within groups than between. But precise definitions of what constitutes a community differ depending on context and assumptions, resulting in a manifold of justifiable characterisations (Fortunato 2010). Classical centrality measures neglect the mesoscopic scale of communities and can often not distinguish between nodes with the same features or nodes embedded in similar network regions. For example, degree centrality assigns the same score to same-degree nodes, and PageRank cannot distinguish between nodes receiving the same amount of support.

To address this issue, network scientists have developed community-aware centrality scores that typically define node importance in terms of intra-community and inter-community link patterns. They commonly evaluate their effectiveness in a disease spreading setting where the objective is to contain an epidemic by immunising a limited fraction of the population (Cherifi et al. 2019; Masuda 2009; Ghalmane et al. 2019; Rajeh et al. 2021). For example, community-based betweenness centrality considers only shortest paths with endpoints in different communities (Kitromilidis and Evans 2018). Community hub-bridge calculates a node's importance as the sum of its intra-community and inter-community links, weighted by the size of the node's community and the number of communities it connects to, respectively (Ghalmane et al. 2019). Community-based centrality determines a node's importance as the number of connections it has to other communities, weighted by the communities' relative sizes (Zhao et al. 2015). Masuda proposed a measure based on eigenvector centrality that quantifies a node's centrality in terms of its contribution to the connectivity between modules, giving higher importance to those nodes that, if removed, would fragment the network more (Masuda 2009). Modular centrality defines a generic framework that operates on top of classical centrality measures to retrofit them with community awareness. It decomposes the network into local, intra-community, and global, inter-community parts and represents a node's centrality as a combination of a local and global component (Ghalmane et al. 2019). In networks with overlapping community structures, nodes that belong to several communities may have high influence despite having a low degree because they act as bridges between communities (Kumar et al. 2018). For epidemic settings, a node's number of community memberships, sometimes called membership centrality, is typically at least as good an estimator of influence as global centrality measures (Hébert-Dufresne et al. 2013). Assuming a network's overlapping community structure is known, random walk-based approaches can be employed to extract high-degree nodes from overlapping regions (Taghavian et al. 2017). Overlapping modular centrality generalises modular centrality and takes into account the possibly multiple community memberships that nodes have, resulting in increased influence in the local parts of a network (Ghalmane

et al. 2019). Recently, modularity vitality has been proposed (Magelinski et al. 2021) based on the community-detection approach known as modularity (Newman and Girvan 2004) and generalised to overlapping communities (Rajeh et al. 2021).

We focus on non-overlapping community structure and derive a centrality score from the information-theoretic community-detection method known as the map equation (Rosvall and Bergstrom 2008) analytically: *map equation centrality*. Deriving community-aware centrality scores from a community-detection approach provides more clarity and precision because the resulting measures adopt the same assumptions regarding what constitutes communities as the underlying community-detection approach. The map equation framework uses random walks to model network flows and identifies communities as those network regions where a random walker tends to spend a long time before switching to a different region. Therefore, map equation centrality determines node importance from a flow-based perspective. Using a toy example, we highlight how map equation centrality exploits community structure to distinguish between nodes where classical centrality measures fail. To understand how map equation centrality is affected by randomness in the link patterns of a network, we generate an Lancichinetti-Fortunato-Radicchi (LFR) network with planted community structure, rewire different fractions of the links, and compare the resulting community structures and node centralities with the ground truth. To evaluate the performance of map equation centrality, we apply it to twelve empirical networks to identify influential nodes. Like in previous work on centrality scores, we contrast our predictions with the spreading power of nodes obtained from simulations of a Susceptible-Infected-Recovered (SIR) disease-spreading model (Ghalmane et al. 2019; Rajeh et al. 2021) and the adoptions of ideas modelled by the linear threshold model (Rajeh et al. 2022). For comparison, we include degree centrality as a local measure, betweenness centrality as a global measure, as well as three other community-aware centrality measures in our evaluation. We find that map equation centrality performs amongst the best in half of the networks in the SIR setting and tends to outperform the baseline measures in the linear threshold setting.

The map equation framework

The map equation (Rosvall and Bergstrom 2008) is a flow-based information-theoretic objective function for community detection. It takes a network $G = (V, E, \delta)$, possibly weighted and/or directed, and a partition M of the network's nodes into modules as input, and measures how well the partition captures the network's community structure. Here, V is the set of nodes, $E \subseteq V \times V$ is the set of links, and $\delta : E \to \mathbb{R}^+$ is a function that assigns weights to the links. A partition M is a split of the network's nodes into disjoint, possibly nested sets.

Conceptually, the map equation models network flow with a random walk on the network and calculates how many bits are required, on average, to encode one random-walker step. To explain the inner workings of the map equation, we consider a communication game where the sender updates the receiver about the location of a random walker on a network. We assume that, when at node u, the probability that the random walker chooses an outgoing link $e = (u, v) \in E$ is proportional to the link's weight, $\delta(e)$.

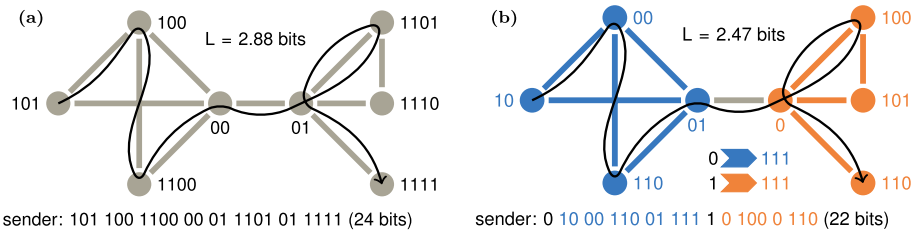

Fig. 1 A communication game on a network where colours indicate module assignments and node labels show codewords. The black trace shows a possible node sequence during a random walk; the corresponding sequence of codewords to describe the walk is shown in the bottom. The average per-step codelength is shown as L. **a** The one-level partition where all nodes are in the same module and there is only one codebook. **b** Nodes are split into two modules with one codebook per module and an additional index-level codebook, indicated by coloured arrows. Module entry and exit codewords are shown on the left and right of the arrows, respectively. The codelength L is reduced because the partition captures those areas where the random walker tends to stay for a longer time

In the simplest case, when there is only one module that contains all nodes, we assign unique codewords to the nodes according to a Huffman code based on the nodes' visit rates at ergodicity. We refer to such a partition as the one-level partition and denote it as M_1. When the random walker takes a step, the sender communicates one codeword to the receiver (Fig. 1a). According to Shannon's source coding theorem (Shannon 1948), the lower bound for the per-step codelength, L, is precisely the entropy of the nodes' visit rates,

$$L(G, \mathsf{M}_1) = H(P) = -\sum_{u \in V} p_u \log_2 p_u, \tag{1}$$

where H is the Shannon entropy, P is the set of node visit rates, and p_u is the visit rate of node u.

In undirected networks, we calculate the node visit rates analytically as $p_u = \frac{s_u}{\sum_{v \in V} s_v}$, where $s_u = \sum_{v \in V} \delta((u, v))$ is the strength of node u. In directed networks, we obtain the visit rates numerically as the stationary distribution of a random walk on the network. The Perron-Frobenius theorem guarantees the existence of such an ergodic distribution in strongly connected networks; to ensure ergodicity in weakly-connected networks, there are different options. PageRank relaxes these dynamics by introducing uniform node teleportation, letting the random walker teleport to a node selected uniformly at random at some small rate (Gleich 2015), introducing a teleportation parameter. To reduce the effect of this parameter, an alternative is so-called unrecorded link teleportation (Lambiotte and Rosvall 2012), a similar approach where the random walker teleports, at some small rate, to links proportionally to their weight.

In networks with community structure, we can achieve shorter codelengths than with the one-level partition. Splitting the nodes into modules allows us to assign unique codewords within modules, and re-use codewords across modules. However, we need to pay for this by encoding transitions between modules: we introduce one designated exit codeword per module, as well as an index-level codebook for encoding transitions into modules. Now, the sender communicates one codeword for transitions within modules, and three codewords for transitions between modules, that is one module exit codeword from the old module codebook, one module entry codeword from the index-level

codebook, and one codeword from the new module codebook to visit a node in the new module (Fig. 1). The codelength for such a two-level map is given by the sum of the index-level entropy and the module-level entropies, weighted by the rate at which each codebook is used,

$$L(G, \mathsf{M}) = qH(Q) + \sum_{\mathsf{m} \in \mathsf{M}} p_{\mathsf{m}} H(P_{\mathsf{m}}). \tag{2}$$

Here, $P_{\mathsf{m}} = \{p_u \mid u \in \mathsf{m}\} \cup \{\mathsf{m}_{\text{exit}}\}$ is the set of node visit rates for module m, including the module exit rate for module m, m_{exit}, and $p_{\mathsf{m}} = \sum_{p \in P_{\mathsf{m}}} p$ is the rate at which the sender uses the codebook for module m. $Q = \{\mathsf{m}_{\text{enter}} \mid \mathsf{m} \in \mathsf{M}\}$ is the set of module entry rates, and $q = \sum_{q_{\mathsf{m}} \in Q} q_{\mathsf{m}}$ is the rate at which the sender uses the index-level codebook.

When a partition reflects the structure of the network well and groups those nodes together where the random walker stays for a longer time, transitions between modules occur at a low frequency, overall compressing the average per-step codelength. Thus, finding the optimal partition according to the map equation becomes a search problem. Through recursion, we can generalise this approach to partitions nested at arbitrary depth and reduce the codelength even further in networks with hierarchical community structure.

Map equation centrality

To define our community-aware centrality score, map equation centrality, we take inspiration from the concept of network vitality. Given a function f that operates on networks and calculates a numerical value, the vitality $\mu(G, u)$ with respect to a node u is defined as

$$\mu(G, u) = f(G) - f(G - \{u\}), \tag{3}$$

where $G - \{u\}$ denotes G with u removed (Koschützki et al. 2005). But because removing a node and its incident links from the network would disrupt the network's community structure and change the nodes' visit rates, instead, we keep the network unchanged and only omit u when describing the community structure—we call this *silencing a node*. We realise silencing with the Vickrey–Clarke–Groves (VCG) principle for setting prices in multi-item auctions, such as AdWords auctions, a generalisation of second-price sealed-bid auctions for single items where the bidder who submits the highest bid for an item receives the item for the value of the second-highest bid (Vickrey 1961). The VCG mechanism determines the price that bidder b has to pay for item i as the marginal harm caused to other bidders who, because of b's existence, receive an item $j \neq i$ that they value lower than i. The price that b pays for i is "the difference between the optimal valuation achievable by allocating everyone except person b to all the positions and the optimal valuation obtainable by allocating everyone except person b to all positions other than i" (Leonard 1983). Specifically, b's price for i does not depend on b's own wealth but is determined by the collective marginal harm caused to the remaining bidders. Following the same idea, we define a node u's importance as the collective marginal harm it causes to the remaining nodes in terms of codeword length, that is, by how many bits the codeword lengths for the remaining nodes could be reduced if u was silenced.

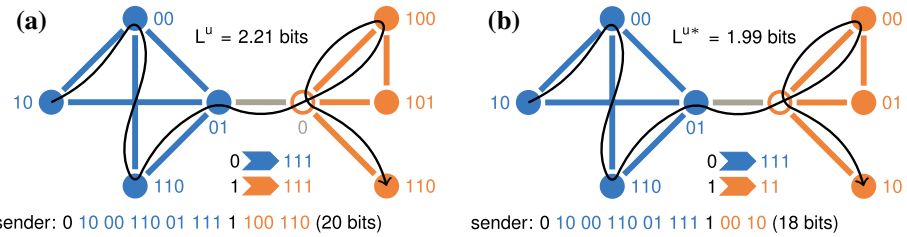

Fig. 2 Two options for describing a random walk when a node is silenced, with the silenced node shown as a ring. In both cases, the sender still communicates module entries through the silenced node. **a** Using the same code as before: when the random walker visits the silenced node, the sender does not use the corresponding node-visit codeword. **b** Designing a new code: the silenced node does not receive a codeword and visits to that node cannot be encoded

In terms of the communication game, silencing a node means that, when the random walker visits a silenced node u, the sender does not communicate the codeword for visiting u to the receiver (Fig. 2a). But this is inefficient because node u has a codeword that is never used, that is, the sender uses more bits than necessary to describe the random walk. Instead, we can design a new coding scheme without assigning a codeword to u and, thereby, compress the description of the random walk (Fig. 2b). Map equation centrality is always positive because silencing a node deletes its codeword from the coding scheme, making it possible to assign shorter codewords to the remaining nodes. Following the VCG principle, we define the centrality of node u as the difference between the original, inefficient code—we call it L^u—and the updated, efficient code—we call it L^{u*},

$$\lambda(G, \mathsf{M}, u) = L^u(G, \mathsf{M}) - L^{u*}(G, \mathsf{M}). \tag{4}$$

Paraphrasing the VCG principle, map equation centrality for node u is the codelength difference between the optimal coding scheme that assigns codewords to all nodes but never uses the codeword for node u and the optimal coding scheme that assigns codewords to all nodes but u. We derive expressions for L^u and L^{u*} from the map equation, and, for clarity, begin with one-level partitions, then moving on to two-level and hierarchical partitions.

First, we consider the case where we use the old coding scheme. We obtain the codelength resulting from silencing u from Eq. 1 by removing u from the summation,

$$L^u(G, \mathsf{M}_1) = -\sum_{v \in V, v \neq u} p_v \log_2 p_v. \tag{5}$$

Designing a new coding scheme without a codeword for u changes the codeword lengths for the rest of the nodes. Before, the codeword length for some node v was given by its visit rate as $\log_2 p_v$, but now that u does not receive a codeword anymore, we need to renormalise accordingly. The new codeword length for node $v \neq u$ is $\log_2 \frac{p_v}{1-p_u}$, and for u it is zero, resulting in a codelength of

$$L^{u*}(G, \mathsf{M}_1) = -\sum_{v \in V, v \neq u} p_v \log_2 \frac{p_v}{1 - p_u}. \tag{6}$$

Plugging Eqs. 5 and 6 into Eq. 4, we get u's contribution to the codelength in the one-level partition M_1,

$$\begin{aligned}\lambda(G, \mathsf{M}_1, u) &= L^u(G, \mathsf{M}_1) - L^{u*}(G, \mathsf{M}_1) \\ &= -(1 - p_u)\log_2(1 - p_u).\end{aligned} \tag{7}$$

We move on to derive the same quantities for two-level partitions M. Again, we begin by considering the resulting codelength when silencing node u but using the old coding scheme. Then, we design a new coding scheme that does not assigning a codeword to u, and calculate the difference between the two coding schemes to obtain u's contribution. For clearer derivations, we distinguish explicitly between m_u, the module that contains u, and the rest of the modules by rewriting the map equation (Eq. 2),

$$L(G, \mathsf{M}) = qH(Q) + \sum_{\mathsf{m}\in\mathsf{M}} p_{\mathsf{m}} H(P_{\mathsf{m}}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\mathsf{m}\in\mathsf{M},\mathsf{m}\neq\mathsf{m}_u} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{modules without } u} - \overbrace{\sum_{p\in P_{\mathsf{m}_u}} p\log_2\frac{p}{p_{\mathsf{m}_u}}}^{\text{module with } u}. \tag{8}$$

From Eq. 8, it becomes clear that silencing node u in a two-level partition only affects the module that contains u because a codeword for u only exists in the context of m_u, but not in other modules. The codelength for a two-level partition M, using the old coding scheme while u is silenced is

$$L^u(G, \mathsf{M}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\mathsf{m}\in\mathsf{M},\mathsf{m}\neq\mathsf{m}_u} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{modules without } u} - \overbrace{\sum_{p\in P_{\mathsf{m}_u}\setminus\{p_u\}} p\log_2\frac{p}{p_{\mathsf{m}_u}}}^{\text{module with } u}. \tag{9}$$

Because of the modular structure of the coding scheme, when designing a new code, only codewords for nodes in module m_u are affected while other modules and the index level remain unaffected. The new codebook usage rate for module m_u is $p_{\mathsf{m}_u} - p_u$, which is also the term we use for re-normalising the node visit rates for nodes in m_u. That is, the new rate at which the codeword for $v \in \mathsf{m}_u$ with $v \neq u$ is used is $\frac{p_v}{p_{\mathsf{m}_u} - p_u}$, and the module exit codeword is used at rate $\frac{\mathsf{m}_u\text{exit}}{p_{\mathsf{m}_u} - p_u}$. The new codelength for M is

$$L^{u*}(G, \mathsf{M}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\mathsf{m}\in\mathsf{M},\mathsf{m}\neq\mathsf{m}_u} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{modules without } u} - \overbrace{\sum_{p\in P_{\mathsf{m}_u}\setminus\{p_u\}} p\log_2\frac{p}{p_{\mathsf{m}_u} - p_u}}^{\text{module with } u}. \tag{10}$$

Plugging Eqs. 9 and 10 into Eq. 4, we get u's contribution to the two-level codelength where the terms for the index level and those modules that do not contain u cancel out,

$$\begin{aligned}\lambda(G, \mathsf{M}, u) &= L^u(G, \mathsf{M}) - L^{u*}(G, \mathsf{M}) \\ &= -\sum_{p\in P_{\mathsf{m}_u}\setminus p_u} p\log_2\frac{p_{\mathsf{m}_u} - p_u}{p_{\mathsf{m}_u}} \\ &= -\left(p_{\mathsf{m}_u} - p_u\right)\log_2\frac{p_{\mathsf{m}_u} - p_u}{p_{\mathsf{m}_u}}\end{aligned} \tag{11}$$

For the one-level partition M_1, the expression in Eq. 11 reduces to Eq. 7 because all nodes are in the same module and, consequently, $p_{\mathsf{m}_u} = 1$.

Through recursion, we can extended map equation centrality to hierarchical partitions with more than two levels. In fact, since silencing a node u only affects module m_u, Eq. 11 can be used to calculate centralities for nodes in modules that are nested deeper

in the module hierarchy of a network. Further, we can extend map equation centrality to silencing a set of nodes by adjusting Eqs. 9 and 10 (see "Appendix"), leading to

$$\lambda(G, \mathsf{M}, U) = -\sum_{\mathsf{m}\in\mathsf{M}, \mathsf{m}\cap U\neq\emptyset} (p_{\mathsf{m}} - p_{\mathsf{m}\cap U}) \log_2 \frac{p_{\mathsf{m}} - p_{\mathsf{m}\cap U}}{p_{\mathsf{m}}}. \tag{12}$$

Here, U is the set of nodes that are silenced, and $p_{\mathsf{m}\cap U} = \sum_{u\in\mathsf{m}\cap U} p_u$ is the sum of visit rates for the silenced nodes in module m. Moreover, map equation centrality is agnostic to the chosen flow model and can be used with standard PageRank, unrecorded link teleportation, or other suitable flow models that may be chosen based on the dynamic process that is analysed. Map equation centrality can be generalised to overlapping communities through memory networks (Edler et al. 2017) using trajectory data to determine link weights.

Map equation centrality relates to the Kullback-Leibler divergence, also known as relative entropy, and defined as $D_{KL}(P||Q) = -\sum_{x\in X} p(x) \log_2 \frac{q(x)}{p(x)}$, where X is a set of events, and P and Q are probability distributions over X. The KL divergence quantifies the expected number of extra bits that are required to encode a sequence of events with true distribution P, assuming that we use a code optimised for Q. In this light, the importance of a node u is the Kullback-Leibler divergence between encoding visits in module m_u with true codebook usage rate p_{m_u} and silencing u, resulting in a new codebook usage rate after silencing of $p_{\mathsf{m}_u} - p_u$. Because no other modules than m_u contribute to our score, u's importance under map equation centrality is fully determined by its own visit rate p_u and its modular context through p_{m_u}.

Application to synthetic and empirical networks

We have implemented map equation centrality in Infomap, a fast and greedy optimisation algorithm for the map equation with an open source implementation available on GitHub[1] (Edler et al. 2020). In a network with n nodes, Infomap detects communities and computes codeword usage rates for all nodes and codebook usage rates for all modules in time $\mathcal{O}(n \log n)$ (Edler et al. 2017). With this information available, traversing the network partition and computing map equation centrality scores for all n nodes takes time $\mathcal{O}(n)$. Detecting communities and computing map equation centrality scores combined takes time $\mathcal{O}(n \log n)$.

To evaluate map equation centrality, we apply it to synthetic and empirical networks. First, using a toy example, we highlight how map equation centrality overcomes traditional centrality scores' inability to distinguish between same-feature nodes when adopting a local or global point of view. Second, we generate an LFR network with strong community structure and measure how the ranking of nodes according to map equation centrality changes as we rewire different fractions of the network's links. Third, we evaluate map equation centrality alongside two traditional and three community-aware centrality scores on a set of empirical social, biological, web, co-authorship, and infrastructure networks using two different spreading processes, (i) the linear threshold model and (ii) the Susceptible-Infected-Recovered (SIR) disease spreading model.

[1] https://github.com/mapequation/infomap.

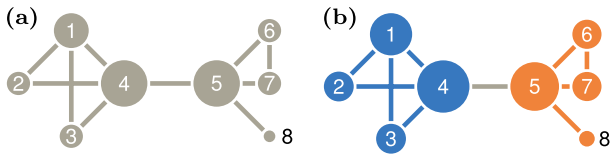

Fig. 3 Illustration of the centrality scores from Table 1. Node colours indicate community assignments, node diameter is proportional to **a** degree centrality and PageRank without teleportation, and **b** map equation centrality

Table 1 Rounded centrality scores for the toy network: degree centrality (DC), betweenness centrality (BC), PageRank without teleportation (PR), and map equation centrality (λ) for the one-level partition M_1, a sub-optimal partition M_{sub}, and the optimal partition M_{opt}, shown in (Fig. 3)

u	DC	BC	PR	$\lambda(\mathsf{M}_1)$	$\lambda(\mathsf{M}_{\text{sub}})$	$\lambda(\mathsf{M}_{\text{opt}})$
1	0.43	0.02	0.15	0.20	0.16	0.184
2	0.29	0	0.10	0.14	0.12	0.130
3	0.29	0	0.10	0.14	0.12	0.130
4	0.57	0.60	0.20	0.26	0.24	0.228
5	0.57	0.67	0.20	0.26	0.24	0.212
6	0.29	0	0.10	0.14	0.13	0.127
7	0.29	0	0.10	0.14	0.13	0.127
8	0.14	0	0.05	0.07	0.07	0.068

We explore the centrality scores through the lens of two different spreading processes because they highlight various aspects. Neither of them is more valid than the other but they are simply tools for comparison of different use cases. In both cases, we test two different flow models as a basis for community detection with Infomap: (a) unrecorded link teleportation (Lambiotte and Rosvall 2012), and (b) recorded node teleportation, corresponding to standard PageRank with teleportation rate 0.15 (Gleich 2015). In principle, one could define further domain-specific flow models, determine node visit rates through simulations, and use them as an input for Infomap. For reproducibility, we provide our code for evaluation in a GitHub repository.[2]

Toy example: how map equation centrality discerns same-feature nodes

We use a small, undirected network with eight nodes and ten links (Fig. 3), and use `networkx` (Hagberg et al. 2008) and Infomap to calculate centrality scores for its nodes (Table 1). The optimal way to partition the network, by design and recovered by Infomap, is to group the nodes into two communities as indicated by colours (Fig. 3b).

We find that neither degree centrality nor map equation centrality when based on the one-level partition $\mathsf{M}_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ can distinguish between nodes with the same degree (Fig. 3a, Table 1). This is because using M_1 turns map equation centrality into a global approach, ignoring the network's mesoscopic community structure. However, when using the sub-optimal two-level partition $\mathsf{M}_{\text{sub}} = \{\{1, 2, 3\}, \{4, 5, 6, 7, 8\}\}$ with codelength 3.24 bits, or the optimal two-level partition $\mathsf{M}_{\text{opt}} = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}\}$

[2] https://github.com/mapequation/map-equation-centrality.

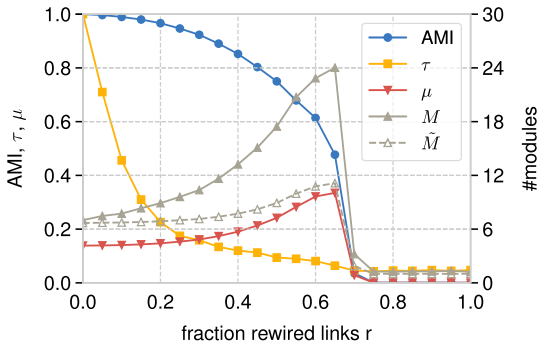

Fig. 4 Results under rewiring of an LFR network. For each fraction of rewired links, *r*, we infer the community structure with Infomap and compute map equation centrality scores. We report AMI with the ground truth partition, correlation with the node ranking under the ground truth partition, τ, the mixing, μ, as well as the number of detected communities M and effective number of communities $\tilde{M}$. The reported values are averages over 100 rewirings for each *r*

with codelength 2.47 bits (Fig. 3b), map equation centrality distinguishes between same-degree nodes that are embedded in different modules while same-degree nodes in the same module remain indistinguishable (Fig. 3b, Table 1). We explain this by interpreting Eq. 11: the importance of a node u is determined by its visit rate, p_u, as well as the codebook usage rate of its module, p_{m_u}, that is, modules with a higher codebook usage rate boost the importance of their member nodes to a higher degree than modules with a lower codebook usage rate.

Synthetic network: behaviour of map equation centrality under link rewiring

We generate an LFR network (Lancichinetti et al. 2008) with 1000 nodes, average degree $k = 10$, minimum community size 100, node degree exponent $\gamma = 2.5$, community size exponent $\beta = 1.5$, and mixing parameter $\mu = 0.1$. The resulting network has 7 communities, and, using those communities, we calculate map equation centrality scores for all nodes. We then rewire an r-fraction of the network's links and use Infomap to detect communities M in the rewired network and Kendall's τ coefficient to measure how the nodes' ranking has changed. With adjusted mutual information (AMI), we estimate the agreement between the new communities and the ground truth community structure. We also compute the effective number of communities as the perplexity over the relative modules' sizes, $\tilde{M} = 2^{H(\mathsf{M})}$, where $H(\mathsf{M}) = \sum_{\mathsf{m}\in\mathsf{M}} \frac{|\mathsf{m}|}{N} \log_2 \frac{|\mathsf{m}|}{N}$ is the Shannon entropy of the relative module sizes, N is the number of nodes in the networks, and $|\mathsf{m}|$ is the number of nodes in module m. The effective number of modules is the number of same-size modules with the same entropy into which the nodes would be partitioned. An effective number of modules close to the actual number of detected modules indicates that the detected communities have similar size, whereas a much smaller number of effective modules indicates partition with a smaller number of large modules and a larger number of small modules. For robust results, we repeat the rewiring for each r 100 times and report average values for AMI, τ, resulting mixing μ, the number of communities M, and the number of effective communities $\tilde{M}$; the results are shown in Fig. 4.

Overall, we see that small amounts of noise caused by rewiring can affect the node ranking to a larger extent despite a relatively stable number of communities with high AMI values.

Datasets and methods

We use twelve real-world networks, retrieved from netzschleuder (Peixoto 2020), to evaluate map equation centrality's performance. Seven of the networks are undirected while five are directed.

Facebook friends undirected network of Facebook friendships, recorded in April 2014, where a link between users A and B means that they are friends on Facebook (Maier and Brockmann 2017).

Copenhagen undirected network of Facebook friendships between university students from Copenhagen where a link between users A and B means that they are friends on Facebook (Sapiezynski et al. 2019).

Uni email directed network of email exchanges at the Rovira i Virgili University in Spain, recorded in 2003, where a link from user A to user B means that user A has sent an email to user B (Guimerà et al. 2003).

Polblogs directed network of U.S. political blog websites, recorded in 2004, where a link from blog A to B means that A has a hyperlink to B (Adamic and Glance 2005).

Interactome yeast undirected network of yeast proteins where a link between proteins A and B means that they interact with each other (Coulomb et al. 2005).

Ego Facebook undirected network of Facebook friendships, recorded in 2012, where a link between users A and B means they are friends on Facebook (Mcauley and Leskovec 2014).

Power undirected network of the power grid in the western U.S. where nodes represent generators, transformers, and substations, and they are connected by a link if a high-voltage transmission line runs between them (Watts and Strogatz 1998).

Facebook organizations undirected network of Facebook friendships between users working at the same organization, a link between users A and B means that they are friends on Facebook (Fire and Puzis 2016).

Physics collaborations undirected co-authorship network between researchers who have a preprint on arXiv, recorded in May 2014, where a link between researcher A and B means that they have written an arXiv preprint together (De Domenico et al. 2015).

Google directed network of hyperlinks between internal websites at Google, recorded in 2004. A link from page A to B means that there is a hyperlink from A to B (Palla et al. 2007).

PGP directed network of users in the Pretty-Good-Privacy (PGP) web of trust, recorded in November 2009. A link from user A to user B means that user A trusts user B (Richters and Peixoto 2011).

Facebook wall directed network of interactions between Facebook users, recorded in 2009, where a link from user A to user B means that user A has posted on user B's wall (Viswanath et al. 2009).

Table 2 Details for eight empirical networks: their number of nodes, N, number of links, $|E|$, average degree, k, epidemic threshold, p_{th}; the number of communities inferred with Infomap, M, the effective number of communities $\tilde{M}$, and mixing, μ, both for link and node teleportation

Network	N	$\|E\|$	k	p_{th}	Link teleportation			Node teleportation		
					M	$\tilde{M}$	μ	M	$\tilde{M}$	μ
Facebook friends	329	1954	11.9	0.048	21	13	0.129	22	13	0.115
Copenhagen	800	6429	16.1	0.038	37	29	0.499	36	29	0.499
Uni email*	1133	5452	19.2	0.027	50	32	0.406	55	34	0.409
Polblogs*	1222	19,024	31.1	0.010	88	6	0.164	51	6	0.177
Interactome yeast	1458	1993	2.7	0.161	166	142	0.237	178	154	0.247
Ego Facebook	4039	88,234	43.7	0.009	77	32	0.082	86	35	0.083
Power	4941	6594	2.7	0.348	428	377	0.163	465	416	0.177
Facebook organizations	5524	94,219	34.1	0.016	53	30	0.360	62	40	0.391
Physics collaborations	8798	27,416	6.2	0.066	610	491	0.218	656	537	0.227
Google*	15,763	171,206	21.7	0.001	597	260	0.470	600	225	0.518
PGP*	39,796	301,498	15.2	0.010	2851	1529	0.285	3285	1843	0.305
Facebook wall*	43,953	271,375	12.3	0.028	2995	1375	0.493	3228	1747	0.519

Directed networks are marked with *

Table 2 provides details about the networks' size, average node degree, epidemic threshold; their number of communities as detected with Infomap, effective number of communities, and mixing, both for link and node teleportation. Since estimating nodes' spreading power with the SIR simulation as well as the linear threshold model disregard link weights, we treat all networks as unweighted.

To infer the networks' community structure, we select the solution with the shortest codelength from 1000 Infomap runs, both using unrecorded link teleportation and recorded node teleportation with teleportation rate 0.15 where the latter corresponds to standard PageRank. We test different flow models because they describe different dynamic processes on the network, lead to different community structures, and are therefore suitable for different applications. In our evaluation, we consider two-level partitions with non-overlapping communities. We have also tested hierarchical partitions, but did not see a substantial performance difference. For comparison, we include degree centrality as a local measure, betweenness centrality as a global measure, and the three community-aware centrality scores modularity vitality (Magelinski et al. 2021), community hub-bridge (Ghalmane et al. 2019), and community-based centrality (Zhao et al. 2015). Modularity vitality calculates a node u's importance, given a network G and a partition M, as the difference in modularity between the original network and partition and the network and partition with u removed, $Q(G, \mathsf{M}) - Q(G - \{u\}, \mathsf{M} - \{u\})$, where Q is the modularity function. Depending on whether deleting a node and its incident links increases or decreases the partitions modularity, the result can be positive or negative. Following previous evaluations, we consider modularity vitality's absolute value (Rajeh et al. 2021). Community hub-bridge determines a node u's importance by considering its intra- and inter-community links, weighing them by u's own community size and the number of other communities it links to, respectively, assigning high importance to nodes with many links in large communities and nodes with many links to a large number

of communities, $\sum_{\mathsf{m}\in\mathsf{M}} |\mathsf{m}| \cdot k_u^{\mathsf{m}} + \mathrm{NNC}_u \cdot k_u^{\overline{\mathsf{m}}}$. Here, k_u^{m} is the number of u's neighbours in module m, NNC_u is u's number of neighbouring communities, and $k_u^{\overline{\mathsf{m}}}$ is the number of u's neighbours outside of m. Community-based centrality calculates a node's importance as the number of connections it has to the different communities, weighted by the communities' relative sizes, $\sum_{\mathsf{m}\in\mathsf{M}} k_u^{\mathsf{m}} \frac{|\mathsf{m}|}{N}$.

Evaluation with the linear threshold model

The linear threshold model simulates the spread and adoption of ideas and behaviours through a network and has previously been applied to evaluate the performance of community-aware centrality scores (Rajeh et al. 2022). In the linear threshold model, nodes can be in either of two states, that is, they can be active or inactive. At the beginning of the simulation, we activate an x-fraction of the nodes, selected as the nodes with the highest centrality according to a centrality measure; all other nodes begin inactive. Then, during each time step of the simulation, the inactive nodes check what fraction of their neighbours is active, and get activated if that fraction is at least as high as a given threshold t. This threshold can be uniform across all nodes, or it can be node-dependent. Here, in absence of node-dependent threshold information in the data, we use the uniform threshold $t = 0.5$, and include further results for thresholds $t' = 0.4$ and $t'' = 0.6$ in the "Appendix". The simulation continues until no more nodes get activated; then we count the influence of the initially active nodes in terms of the activation size, that is the fraction of active nodes, where a larger activation size means that the initially active nodes have more influence.

We find that, for large enough fractions of initially active nodes, map equation centrality outperforms the other measures when using the recorded link teleportation-based flow model in most cases, but has lower performance when based on unrecorded link teleportation. Modularity vitality tends to outperform community hub-bridge, and community-based centrality, but is itself often outperformed by betweenness centrality, especially in the larger and directed networks. For small fractions of initially active nodes, community hub-bridge and community-based centrality tend to perform better than map equation centrality and modularity vitality. Further, the flow model choice has a larger effect on map equation centrality than on modularity vitality, community hub-bridge, and community-based centrality. All of the measures tend to perform better when using the PageRank-based communities (Fig. 5a–l). We describe the results in more detail in the "Appendix".

Evaluation with the SIR disease spreading model

The second spreading process we use to evaluate map equation centrality's performance is a discrete-time SIR disease spreading simulation. We follow the approach taken by Rajeh et al. (2021) to test how accurately the centrality measures identify influential nodes.

To estimate a node u's influence, we calculate its spreading power, that is the expected number of nodes that get infected by a disease with the single initial spreader u: Initially, only node u is infected, all other nodes begin in the susceptible state, and the recovery

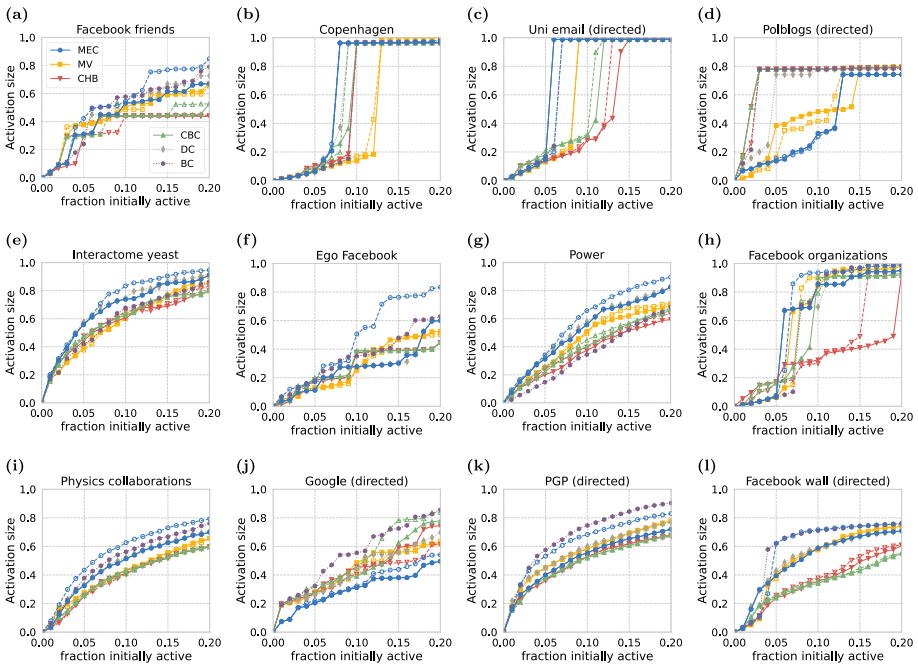

Fig. 5 Activation size for map equation centrality (MEC), modularity vitality (MV), community hub-bridge (CHB), community-based centrality (CBC), degree centrality (DC), and betweenness centrality (BC) in twelve empirical networks under the linear threshold model with threshold $t = 0.5$. Community structures are identified with Infomap; solid lines use the unrecorded link teleportation flow model, dashed lines use recorded node teleportation

time is set to 1 time step. As long as there are infected nodes, the simulation continues. Infected nodes infect their susceptible neighbours independently with probability p_{th}, then they recover. Here, p_{th} is the so-called epidemic threshold (Table 2) with $p_{\text{th}} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$ where $\langle k \rangle = \frac{1}{|V|} \sum_{v \in V} k_v$ and $\langle k^2 \rangle = \frac{1}{|V|} \sum_{v \in V} k_v^2$ are the first and second moment of the network's degree sequence, respectively (Wang et al. 2016). When no infected nodes are left, the simulation ends, and we determine u's spreading power as the number of recovered nodes. Because of the stochasticity in the SIR model, we repeat the simulation 1000 times per node to calculate its expected spreading power.

Let M_c and M_{SIR} be the lists of nodes, ranked according to centrality score c, and their spreading power as determined with the SIR simulation, respectively. Then, we measure the ability of centrality score c to identify influential spreaders using the so-called imprecision function, $\epsilon_c(x) = 1 - \frac{M_c(x)}{M_{\text{SIR}}(x)}$ (Kitsak et al. 2010). Here, $M_c(x)$ and $M_{\text{SIR}}(x)$ are the average spreading power of the top x-fraction of nodes according to centrality score c and the SIR simulation, respectively. A smaller imprecision value corresponds to a better alignment between centrality score c and spreading power.

In four of the tested networks, map equation centrality outperforms modularity vitality, community hub-bridge, and community-based centrality, performs second-to-third best in six networks, is worst or second-worst in the remaining two networks when based on unrecorded link teleportation. The performance is often similar to degree centrality because the nodes' visit rates for unrecorded link teleportation are proportional to their degree in undirected networks (Lambiotte and Rosvall 2012). Our community hub-bridge and community-based centrality implementations in directed networks

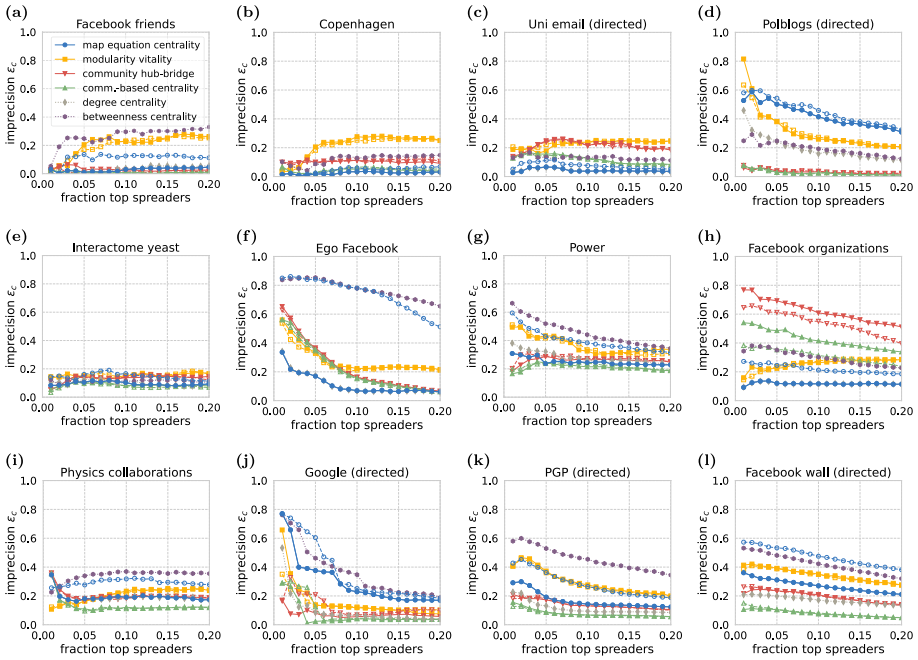

Fig. 6 Imprecision of map equation centrality, modularity vitality, community hub-bridge, community-based centrality, degree centrality, and betweenness centrality for identifying top spreaders in eight empirical networks. The curves show imprecision as a function of the fraction of top spreaders that are selected. A lower imprecision corresponds to more accurately identifying the top spreaders as determined with an SIR simulation. Community structures are identified with Infomap; solid lines use the unrecorded link teleportation flow model, dashed lines use recorded node teleportation. Degree and betweenness centrality do not rely on communities

consider nodes' outgoing links. Because nodes with higher out degrees are expected to infect more nodes in the SIR model, our implementations may explain the measures' good performance. In contrast, map equation centrality cares about the nodes' in-degree because the flow in the map equation framework is based on random walker transitions into the nodes. To calculate degree and betweenness centrality, `networkx` considers the nodes' total degree. With recorded node teleportation, map equation centrality does not perform as well and is often more similar to betweenness centrality. Modularity vitality, community hub-bridge, and community-based centrality are less affected by the choice of flow model (Figs. 6a–l). We describe the results in more detail in the "Appendix".

To investigate whether map equation centrality is at an advantage because it is by definition faithful to the map equation, we have repeated our experiments in the four networks where map equation centrality performed best, using partitions based on modularity maximisation. We infer the community structure in the Copenhagen, Uni email, Ego Facebook, and Facebook organizations networks, with the Louvain algorithm (Blondel et al. 2008), using the `networkx` implementation, and proceed with highest-modularity partitions from 1000 runs with different seeds. Louvain detects 13 (11) communities in the Copenhagen network, 20 (17) communities in the Uni email network, 17 (12) communities in the Ego Facebook network, and 13 (9) communities in the Facebook organizations network, where the numbers in parenthesis are the effective numbers of communities. Overall, we find that map equation centrality with unrecorded

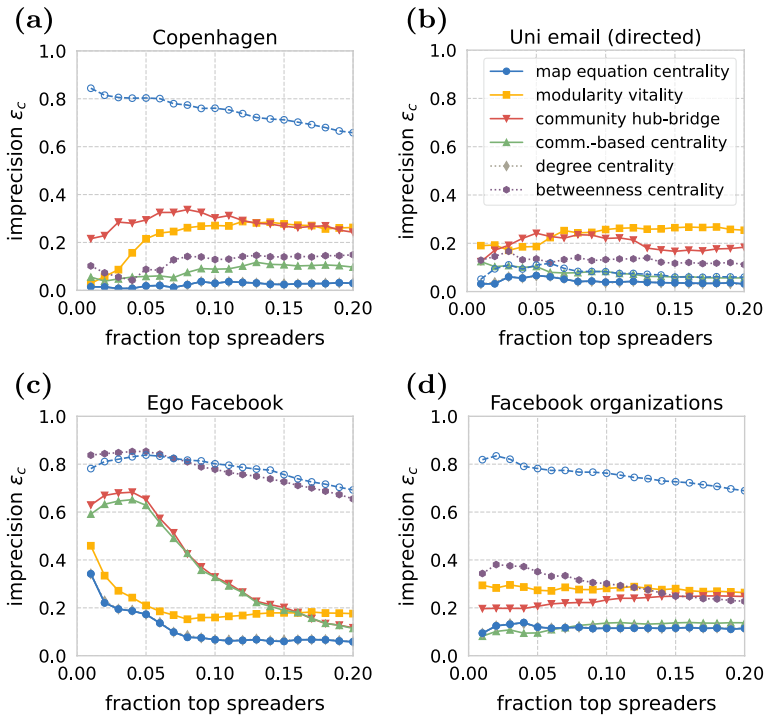

Fig. 7 Imprecision of map equation centrality, modularity vitality, community hub-bridge, community-based centrality, degree centrality, and betweenness centrality in four empirical networks based on partitions inferred thorugh modularity maximisation with the Louvain algorithm

link teleportation and modularity vitality perform similar to before (Fig. 7a–d). Whether community hub-bridge and community-based centrality perform better or worse depends on the network, and map equation centrality with recorded node teleportation performs worse than before.

To summarise, we found that none of the tested centrality scores outperforms all other scores in all networks, but none of the scores performed worst in all cases either.

Distribution of influential nodes

To understand why unrecorded link teleportation facilitates more accurate identification of top spreaders in the SIR case while recorded node teleportation works better for the linear threshold model, we analyse how the top-ranked nodes are distributed across modules. Let M be a partition of the nodes into modules, $\mathsf{m} \in \mathsf{M}$ be a module, and let S be the set of selected nodes by some centrality measure. Then, $\frac{|\mathsf{m} \cap S|}{|S|}$ is the fraction of selected nodes in module m; we calculate the perplexity for S as $2^{H(S)}$ where $H(S) = -\sum_{\mathsf{m} \in \mathsf{M}} \frac{|\mathsf{m} \cap S|}{|S|} \log_2 \frac{|\mathsf{m} \cap S|}{|S|}$. The perplexity corresponds to the effective number of same-size modules across which the selected nodes are distributed uniformly. That is, a higher perplexity means that the selected nodes are more spread out across modules.

The nodes selected by map equation centrality are more spread out with standard PageRank flow than with unrecorded link teleportation. This is because PageRank with a teleportation rate of r assigns an r-fraction of the flow to nodes uniformly, resulting in at least a flow of $\frac{r}{n}$ per node in a network with n nodes. Here, we used $r = 0.15$. For modularity vitality, community hub-bridge, and community-based centrality, the difference

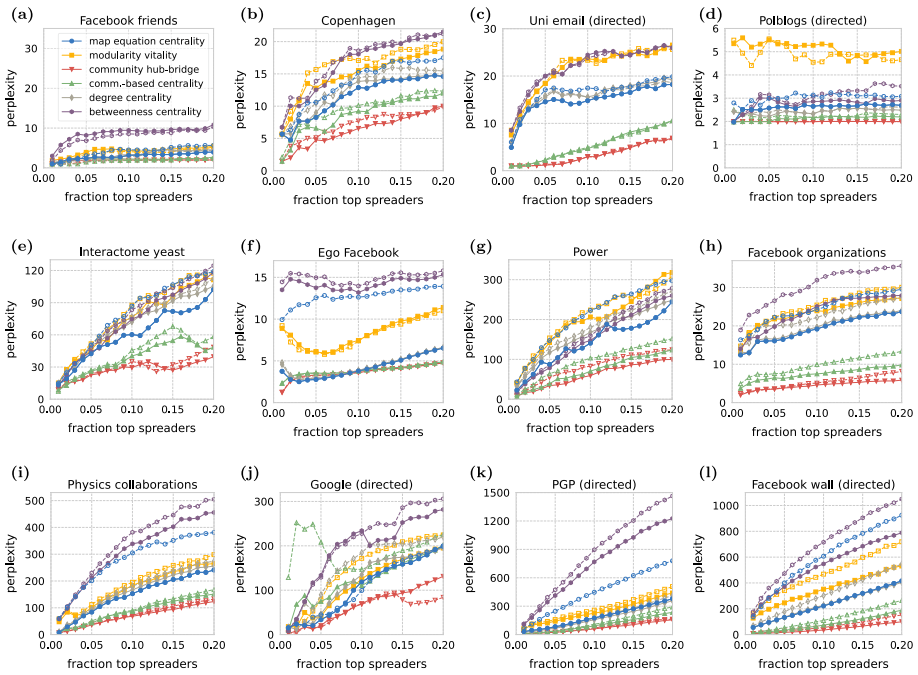

Fig. 8 Perplexity for the distribution of selected nodes as a function of the fraction of selected top spreaders for map equation centrality, modularity vitality, community hub-bridge, community-based centrality, degree centrality, and betweenness centrality in twelve empirical networks. Community structures are identified with Infomap; solid lines use the unrecorded link teleportation flow model, dashed lines use recorded node teleportation

between link and node teleportation is less pronounced, and in some settings even reversed. Overall, community hub-bridge and community-based centrality have lower perplexity, selecting nodes that are less spread out across modules. Map equation centrality and modularity vitality have substantially higher perplexity, spreading out the selected nodes more across communities (Fig. 8a–l). To perform well in the SIR case, a centrality measure should select high-degree nodes because they have a higher opportunity to infect other nodes. Conversely, under the linear threshold model, it is more important to spread out the selected nodes across tightly-knit communities to reach a high activation size, or high-density communities will stop the activation of nodes (Morris 2000).

Conclusion

We have studied node importance from a community-detection perspective within the map equation framework and analytically derived a community-aware centrality score. Our score exploits modular network structure, is agnostic to the chosen flow model, and assigns centrality scores to nodes based on their community embedding; to determine a node's centrality, it suffices to consider those nodes that belong to the same community. In contrast, traditional centrality measures typically neglect local network structure and rely on node features or global patterns to determine node importance instead. Community-aware centrality measures are often defined in an ad-hoc way, disconnected from the assumptions made by community-detection methods. In contrast, map equation centrality is true to the

map equation. We have highlighted how map equation centrality discerns nodes indistinguishable to global centrality measures using a synthetic network. On a set of twelve real-world networks, map equation centrality often performs better than baseline methods in identifying influential nodes.

Appendix

Generalisation for sets of nodes

We generalise map equation centrality and derive the expression in Eq. 12 that can be used to calculate the combined centrality for sets of nodes U. We follow the same approach as before, that is, we first derive an expression for the expected per-step codelength when silencing all nodes in U while using the old coding scheme; then we derive an expression for for the expected per-step codelength when designing a new coding scheme that does not assign codewords to nodes in U to start with.

Let $G = (V, E, \delta)$ be a network with nodes V, links E, weights δ, $U \subseteq V$ be a set of nodes, and $p_U = \sum_{u \in U} p_u$ be the visit rate sum of nodes in U. Further, for a module m, let $p_{\mathsf{m} \cap U} = \sum_{u \in \mathsf{m} \cap U} p_u$ be the visit rate sum of nodes that are members in m and in U, and let $P_{\mathsf{m} \cap U} = \{p_u \mid u \in \mathsf{m} \cap U\}$ be their set of visit rates.

We begin with the one-level partition M_1 and obtain the expected per-step codelength for describing a random walk with nodes in U silenced while using the old coding scheme. Removing the silenced nodes from the summation in Eq. 1, we get

$$L^U(G, \mathsf{M}_1) = -\sum_{v \in V \setminus U} p_v \log_2 p_v. \tag{13}$$

We obtain the codelength for a new coding scheme that does not assign codewords to nodes in U by re-normalising the visit rates for the remaining nodes with $1 - p_U$,

$$L^{U*}(G, \mathsf{M}_1) = -\sum_{v \in V \setminus U} p_v \log_2 \frac{p_v}{1 - p_U}. \tag{14}$$

The difference between Eqs. 13 and 14 is the joint map equation centrality score of the nodes in U under M_1,

$$\begin{aligned} \lambda(G, \mathsf{M}_1, U) &= L^U(G, \mathsf{M}_1) - L^{U*}(G, \mathsf{M}_1) \\ &= -(1 - p_U) \log_2 (1 - p_U). \end{aligned} \tag{15}$$

For two-level partitions, we begin by rewriting the map equation (Eq. 2) to distinguish explicitly between modules that have an overlap with U and those that do not,

$$L(G, \mathsf{M}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\substack{\mathsf{m} \in \mathsf{M} \\ \mathsf{m} \cap U = \emptyset}} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{no overlap with } U} - \overbrace{\sum_{\substack{\mathsf{m} \in \mathsf{M} \\ \mathsf{m} \cap U \neq \emptyset}} \sum_{p \in P_{\mathsf{m}}} p \log_2 \frac{p}{p_{\mathsf{m}}}}^{\text{overlap with } U}. \tag{16}$$

The codelength for describing a random walk in partition M with nodes in U silenced when using the old coding scheme is

$$L^{U}(G,\mathsf{M}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\substack{\mathsf{m}\in\mathsf{M}\\ \mathsf{m}\cap U=\emptyset}} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{no overlap with } U} - \overbrace{\sum_{\substack{\mathsf{m}\in\mathsf{M}\\ \mathsf{m}\cap U\neq\emptyset}} \sum_{p\in P_{\mathsf{m}}\setminus P_{\mathsf{m}\cap U}} p\log_2\frac{p}{p_{\mathsf{m}}}}^{\text{overlap with } U}. \tag{17}$$

With a new code that does not assign codewords to nodes in U and that normalises accordingly, the codelength is

$$L^{U*}(G,\mathsf{M}) = \overbrace{qH(Q)}^{\text{index level}} + \overbrace{\sum_{\substack{\mathsf{m}\in\mathsf{M}\\ \mathsf{m}\cap U=\emptyset}} p_{\mathsf{m}} H(P_{\mathsf{m}})}^{\text{no overlap with } U} - \overbrace{\sum_{\substack{\mathsf{m}\in\mathsf{M}\\ \mathsf{m}\cap S\neq\emptyset}} \sum_{p\in P_{\mathsf{m}}\setminus P_{\mathsf{m}\cap U}} p\log_2\frac{p}{p_{\mathsf{m}}-p_{\mathsf{m}\cap U}}}^{\text{overlap with } U}. \tag{18}$$

The difference between Eqs. 17 and 18 is the joint map equation centrality of the nodes in U under M,

$$\begin{aligned}\lambda(G,\mathsf{M},U) &= L^{U}(G,\mathsf{M}) - L^{U*}(G,\mathsf{M})\\ &= -\sum_{\mathsf{m}\in\mathsf{M},\mathsf{m}\cap U\neq\emptyset} (p_{\mathsf{m}} - p_{\mathsf{m}\cap U})\log_2\frac{p_{\mathsf{m}}-p_{\mathsf{m}\cap U}}{p_{\mathsf{m}}}.\end{aligned} \tag{19}$$

Descriptions of linear threshold model results

In the Facebook friends network, initially all measures perform similarly well. Modularity vitality, community hub-bridge, and community-based centrality outperform map equation centrality between $x = 0.02$ and 0.04; beyond $x = 0.04$, map equation centrality performs best, followed by betweenness centrality, degree centrality, modularity vitality, community-based centrality, and community hub-bridge (Fig. 5a).

In the Copenhagen network, up to $x = 0.03$, all scores perform equally well, between $x = 0.03$ and $x = 0.05$, community hub-bridge and community-based centrality perform slightly better than map equation centrality and modularity vitality. Beyond $x = 0.05$ and up to $x = 0.13$, map equation centrality performs best; for $x \geq 0.13$, modularity vitality performs best and reaches an activation size of 1 (Fig. 5b).

In the Uni email network, initially, community hub-bridge and community-based centrality slightly outperform the other measures. Then, at $x = 0.06$, map equation centrality and degree centrality reach an activation size of nearly 1, followed by betweenness centrality at $x = 0.07$, modularity vitality at $x = 0.09$, community-based centrality at $x = 0.12$, and community hub-bridge at $x = 0.15$ (Fig. 5c).

In the Polblogs network, community-based centrality, community hub-bridge, and betweenness centrality perform best, followed by degree centrality, modularity vitality, and finally map equation centrality (Fig. 5d).

In the Interactome yeast network, map equation centrality performs best, followed by degree centrality, and the remaining measures which have similar performance in this case (Fig. 5e).

In the Ego Facebook network, all four measures have similar performance up to $x = 0.05$, beyond which map equation centrality dominates, followed by betweenness centrality, community-based centrality and community hub-bridge, and modularity vitality (Fig. 5f).

In the Power network, map equation centrality performs best, followed by degree centrality, modularity vitality, community-based centrality, community hub-bridge, and betweenness centrality (Fig. 5g).

In the Facebook organizations network, community hub-bridge and community-based centrality perform best up to $x = 0.05$. Beyond that, map equation centrality performs best; modularity vitality, betweenness centrality, degree centrality, and community-based centrality have similar performance, and community hub-bridge performs weakest with some distance. From $x = 0.14$, betweenness centrality performs slightly better than map equation centrality (Fig. 5h).

In the Physics collaborations map equation centrality outperforms the other measures over the whole tested range, followed by betweenness centrality, degree centrality, modularity vitality, and community hub-bridge and community-based centrality (Fig. 5i).

In the Google network, betweenness centrality performs best while map equation centrality performs weakest in this scenario. The remaining measures have similar performance, but none clearly wins against the others (Fig. 5j).

In the PGP network, betweenness centrality outperforms the remaining measures, followed by map equation centrality, degree centrality, modularity vitality, community hub-bridge, and community-based centrality (Fig. 5k).

Finally, in the Facebook wall network, initially, map equation centrality based on unrecorded link teleportation and degree centrality perform best, followed by community-based centrality, community hub-bridge, betweenness centrality, and modularity vitality. Beyond $x = 0.04$, map map equation centrality with recorded node teleporation and betweenness centrality perform best, followed by modularity vitality, degree centrality, community hub-bridge, and community-based centrality (Fig. 5l).

Descriptions of the SIR model results

In the Facebook friends network, map equation centrality, degree centrality, community hub-bridge, and community-based centrality are nearly tied with an imprecision up to approximately 0.05, identifying the top spreaders accurately. Modularity vitality initially performs similarly well, but achieves imprecision values between around 0.2 and 0.3 beyond $x = 0.05$ (Fig. 6a).

In the Copenhagen network, map equation centrality and degree centrality outperform the other measures. Community-based centrality performs slightly worse than map equation centrality, followed by community hub-bridge, betweenness centrality, and then modularity vitality (Fig. 6b).

In the Uni email network, map equation centrality and degree centrality outperform the other measures across the tested range of x-values, followed by community-based centrality, betweenness centrality, and modularity vitality and community hub-bridge, the latter two performing similarly in this scenario (Fig. 6c).

In the Polblogs network, community-based centrality and community hub-bridge perform best, followed by degree and betweenness centrality, modularity vitality, and finally map equation centrality (Fig. 6d).

In the Interactome yeast network, all measures perform similarly well, while community-based centrality and map equation centrality slightly outperform the rest (Fig. 6e).

In the Ego Facebook network, map equation centrality and degree centrality again outperform the other measures. Initially and up to $x \approx 0.08$, modularity vitality, community hub-bridge, and community-based centrality show similar performance. Beyond $x \approx 0.08$, modularity vitality's performance remains stable at an imprecision of around 0.2 while community hub-bridge and community-based centrality improve and perform as well as map equation centrality at $x \approx 0.2$. Map equation centrality based on recorded node teleportation and betweenness centrality perform substantially worse than the other measures in this scenario with imprecision values roughly between 0.9 down to 0.5 (Fig. 6f).

In the Power network, community-based centrality performs best, followed by map equation centrality, community hub-bridge, degree centrality, modularity vitality, and finally betweenness centrality (Fig. 6g).

In the Facebook organizations network, map equation centrality and degree centrality outperform the other measures with a stable imprecision around 0.1. Modularity vitality performs second-best, with increasing imprecision as x increases, followed by betweenness centrality, community-based centrality, and community hub-bridge (Fig. 6h).

In the Physics collaborations network, modularity vitality initially performs best, but with slightly decreasing performance as x increases. Map equation centrality, degree centrality, community hub-bridge, and community-based centrality initially perform similarly, all with an imprecision of around 0.35, but outperform modularity vitality beyond $x \approx 0.05$, with community-based centrality performing best (Fig. 6i).

In the Google network, up to $x = 0.03$, community hub-bride performs best. Beyond that, community-based centrality performs best, followed by degree centrality, community hub-bridge, modularity vitality, map equation centrality, and finally betweenness centrality (Fig. 6j).

In the PGP network, community-based centrality outperforms the other measures, followed by degree centrality. Community hub-bridge performs third-best, followed by map equation centrality, betweenness centrality, and modularity vitality. In this scenario, node teleportation-based map equation centrality performs nearly identical to modularity vitality. Beyond $x \approx 0.05$, community hub-bridge and map equation centrality are nearly tied (Fig. 6k).

Finally, in the Facebook wall network, community-based centrality outperforms the other measures, followed by degree centrality, community hub-bridge, map equation centrality, modularity vitality, and betweenness centrality. Here, map equation centrality when using recorded node teleportation performs considerably worse compared to unrecorded link teleportation (Fig. 6l).

Further results for the linear threshold model

Further results for the linear threshold model with thresholds $t' = 0.4$ and $t'' = 0.6$ are shown in Figs. 9 and 10, respectively.

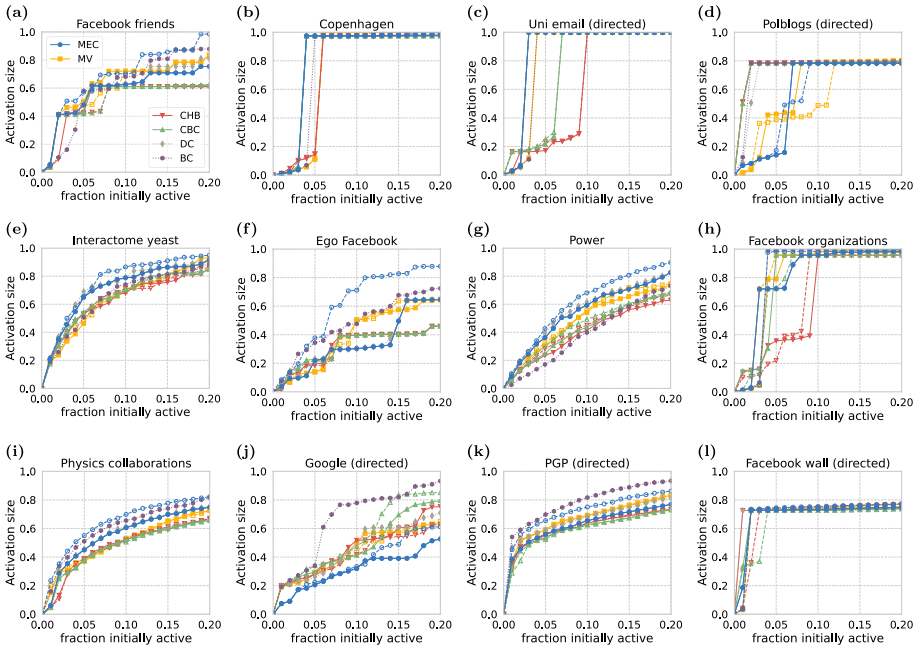

Fig. 9 Activation size for map equation centrality (MEC), modularity vitality (MV), community hub-bridge (CHB), community-based centrality (CBC), degree centrality (DC), and betweenness centrality (BC) in twelve empirical networks under the linear threshold model with threshold $t' = 0.4$. Community structures are identified with Infomap; solid lines use the unrecorded link teleportation flow model, dashed lines use recorded node teleportation

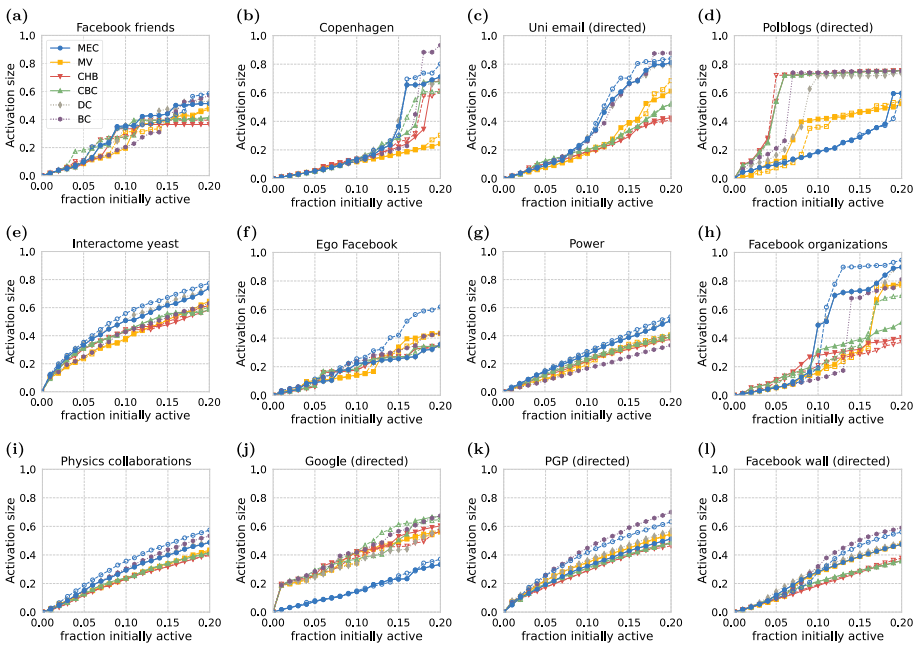

Fig. 10 Activation size for map equation centrality (MEC), modularity vitality (MV), community hub-bridge (CHB), community-based centrality (CBC), degree centrality (DC), and betweenness centrality (BC) in twelve empirical networks under the linear threshold model with threshold $t'' = 0.6$. Community structures are identified with Infomap; solid lines use the unrecorded link teleportation flow model, dashed lines use recorded node teleportation

Acknowledgements
We would like to thank Anton Eriksson for helping with implementing map equation centrality in Infomap, and Jelena Smiljanić and the anonymous reviewers for comments that helped to improve the manuscript.

Author contributions
CB, JCN, and MR conceived and designed the study. CB performed the experiments. CB, JCN, and MR interpreted the results and wrote the paper. All authors read and approved the final manuscript.

Funding
Open access funding provided by Umeå University. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Martin Rosvall was supported by the Swedish Research Council, Grant No. 2016-00796.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the netzschleuder repository, https://networks.skewed.de/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 21 March 2022 Accepted: 26 May 2022
Published online: 16 August 2022

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. election: divided they blog. In: Proceedings of the 3rd international workshop on Link Discovery. LinkKDD '05. Association for Computing Machinery, New York, pp 36–43. https://doi.org/10.1145/1134271.1134277
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008. https://doi.org/10.1088/1742-5468/2008/10/p10008
Cherifi H, Palla G, Szymanski BK, Lu X (2019) On community structure in complex networks: challenges and opportunities. Appl Netw Sci 4(1):117. https://doi.org/10.1007/s41109-019-0238-9
Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat M-C (2005) Gene essentiality and the topology of protein interaction networks. Proc R Soc B Biol Sci 272(1573):1721–1725
De Domenico M, Lancichinetti A, Arenas A, Rosvall M (2015) Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. Phys Rev X 5:011027. https://doi.org/10.1103/PhysRevX.5.011027
Edler D, Bohlin L, Rosvall M (2017) Mapping higher-order network flows in memory and multilayer networks with infomap. Algorithms 10:112. https://doi.org/10.3390/a10040112
Edler D, Eriksson A, Rosvall M (2020) The infomap software package. https://www.mapequation.org
Fire M, Puzis R (2016) Organization mining using online social networks. Netw Spat Econ 16(2):545–578. https://doi.org/10.1007/s11067-015-9288-4
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
Ghalmane Z, El Hassouni M, Cherifi C, Cherifi H (2019) Centrality in modular networks. EPJ Data Sci 8(1):15. https://doi.org/10.1140/epjds/s13688-019-0195-7
Ghalmane Z, Cherifi C, Cherifi H, Hassouni ME (2019) Centrality in complex networks with overlapping community structure. Sci Rep 9(1):10133. https://doi.org/10.1038/s41598-019-46507-y
Ghalmane Z, El Hassouni M, Cherifi H (2019) Immunization of networks with non-overlapping community structure. Soc Netw Anal Min 9(1):1–22. https://doi.org/10.1007/s13278-019-0591-9
Gleich DF (2015) Pagerank beyond the web. SIAM Rev 57(3):321–363. https://doi.org/10.1137/140976649
Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68:065103. https://doi.org/10.1103/PhysRevE.68.06510
Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkX. In: Varoquaux G, Vaught T, Millman J (eds) Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, pp 11–15
Hébert-Dufresne L, Allard A, Young J-G, Dubé LJ (2013) Global efficiency of local immunization on complex networks. Sci Rep 3(1):2171. https://doi.org/10.1038/srep02171
Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43. https://doi.org/10.1007/BF02289026
Kitromilidis M, Evans TS (2018) Community detection with metadata in a network of biographies of western art painters. arXiv:1802.07985

Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888–893. https://doi.org/10.1038/nphys1746
Koschützki D, Lehmann KA, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O (2005) In: Brandes U, Erlebach T (eds) Centrality indices. Springer, Berlin, pp 16–61. https://doi.org/10.1007/978-3-540-31955-9_3
Kumar M, Singh A, Cherifi H (2018) An efficient immunization strategy using overlapping nodes and its neighborhoods. In: Companion proceedings of the the Web Conference 2018. WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp 1269–1275. https://doi.org/10.1145/3184558.3191566
Lambiotte R, Rosvall M (2012) Ranking and clustering of nodes in networks with smart teleportation. Phys Rev E 85:056107. https://doi.org/10.1103/PhysRevE.85.056107
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78:046110. https://doi.org/10.1103/PhysRevE.78.046110
Leonard HB (1983) Elicitation of honest preferences for the assignment of individuals to positions. J Polit Econ 91(3):461–479
Magelinski T, Bartulovic M, M Carley K (2021) Measuring node contribution to community structure with modularity vitality. IEEE Trans Netw Sci Eng 8(1):707–723. https://doi.org/10.1109/TNSE.2020.3049068
Maier BF, Brockmann D (2017) Cover time for random walks on arbitrary complex networks. Phys Rev E 96:042307. https://doi.org/10.1103/PhysRevE.96.042307
Masuda N (2009) Immunization of networks with community structure. New J Phys 11(12):123018
Mcauley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Trans Knowl Discov Data (TKDD) 8(1):1–28. https://doi.org/10.1145/2556612
Morris S (2000) Contagion. Rev Econ Stud 67(1):57–78
Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. https://doi.org/10.1103/PhysRevE.69.026113
Palla G, Farkas IJ, Pollner P, Derényi I, Vicsek T (2007) Directed network modules. New J Phys 9(6):186. https://doi.org/10.1088/1367-2630/9/6/186
Peixoto TP (2020) The Netzschleuder network catalogue and repository. https://networks.skewed.de/
Rajeh S, Savonnet M, Leclercq E, Cherifi H (2021) Comparing community-aware centrality measures in online social networks. In: Computational data and social networks. Springer, Cham, pp 279–290. https://doi.org/10.1007/978-3-030-91434-9_25
Rajeh S, Savonnet M, Leclercq E, Cherifi H (2021) Identifying influential nodes using overlapping modularity vitality. In: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM '21. Association for Computing Machinery, New York, pp 257–264. https://doi.org/10.1145/3487351.3488277
Rajeh S, Yassin A, Jaber A, Cherifi H (2022) Analyzing community-aware centrality measures using the linear threshold model. In: Complex networks & their applications X. Springer, Cham, pp 342–353. https://doi.org/10.1007/978-3-030-93409-5_29
Richters O, Peixoto TP (2011) Trust transitivity in social networks. PLoS ONE 6(4):18384. https://doi.org/10.1371/journal.pone.0018384
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123. https://doi.org/10.1073/pnas.0706851105
Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S (2019) Interaction data from the copenhagen networks study. Sci Data 6(1):1–10. https://doi.org/10.1038/s41597-019-0325-x
Shannon CE (1948) A mathematical theory of communication. Bell Labs Tech J 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Taghavian F, Salehi M, Teimouri M (2017) A local immunization strategy for networks with overlapping community structure. Physica A 467:148–156. https://doi.org/10.1016/j.physa.2016.10.014
Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. J Finance 16(1):8–37
Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in facebook. In: Proceedings of the 2nd ACM workshop on online social networks, pp 37–42 . https://doi.org/10.1145/1592665.1592675
Wang W, Liu Q-H, Zhong L-F, Tang M, Gao H, Stanley HE (2016) Predicting the epidemic threshold of the susceptible-infected-recovered model. Sci Rep 6(1):1–12. https://doi.org/10.1038/srep24676
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442. https://doi.org/10.1038/30918
Zhao Z, Wang X, Zhang W, Zhu Z (2015) A community-based approach to identifying influential spreaders. Entropy 17(4):2228–2252. https://doi.org/10.3390/e17042228

Publisher's Note