Paper review: Finding influential subjects in a network using a causal framework

Lee, Y., Buchanan, A. L., Ogburn, E. L., Friedman, S. R., Halloran, M. E., Katenka, N. V., Wu, J., & Nikolopoulos, G. K. (2023)

Suehyun Kim

3 December 2024

Causal Inference Lab.
Seoul National University

Table of contents

Table of Contents

Introduction

- In a network setting, intervention effects and health outcomes can spill over from one node to another through network ties, and influential subjects are expected to have a greater impact than others.
- Although influence is often defined only implicitly in most of the literature, the operative notion of influence is inherently *causal* in many cases: influential subjects are those we should intervene on to achieve the greatest overall effect across the entire network.
- We define a causal notion of influence using a potential outcome framework and compare it with existing centrality measures, both in terms of assumptions and through simulations.

Table of Contents

Basics of network theory

A **graph** $\mathcal{G} = (V, E)$ is a pair of a set of vertices V and edges E.

- Vertex $i \in \mathbf{V}(\mathcal{G})$: Subjects, nodes, ...
- Edge $e_{ij} \in \mathbf{E}(\mathcal{G})$: Ordered pair of vertices (i, j).
  - Edges can be either *weighted* or *unweighted*.
- Adjacency matrix $\mathbf{E} = [e_{ij}]_{i,j=1}^{N}$: $N \times N$ matrix containing the structural information of a graph. $e_{ij} = 1$ for edges in unweighted graphs.
- Graphs can be either *directed* or *undirected*.
  - $e_{ij} = e_{ji}$ for undirected graphs.

Example. Airport network

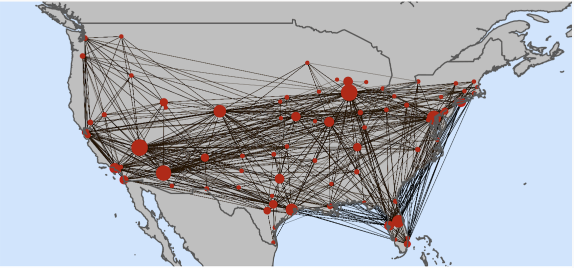


Figure reference: https://jmsallan.netlify.app/

A natural question arises: *Which nodes are important?*

Basics of network theory

Centrality measures: Ways to determine hubs in a graph.

- **Degree centrality**: Degree of a node; every node is as important as other nodes.

$$C_D(u) = \sum_{i=1}^{N} e_{ui}$$

- **Betweenness centrality**: A node is considered important if it works as an efficient 'bridge' between other nodes.

$$C_B(u) = \sum_{i \neq u} \sum_{k \neq i,u} \phi_{ki}(u)/\phi_{ki}$$

  - ϕ_{ki}: # of shortest paths from node k to node i, $\phi_{ki}(j)$: # of shortest paths from node k to node i passing through node j

Basics of network theory

- **Closeness centrality**: Inverse of average shortest path.

$$C_C(u) = \frac{N-1}{\sum_{i=1}^{N} \phi_{ui}}$$

- **Eigenvector centrality**: Component of max eigenvector; a node is important if its neighbors are important.

$$x_u = \frac{1}{\lambda_1} \sum_{i=1}^{N} e_{ui} x_i; \quad \mathbf{E}\mathbf{x} = \lambda_1 \mathbf{x}$$

  - Google PageRank is a variant of eigenvector centrality.

Basics of network theory

Centrality can also be measured based on specific **diffusion processes**.

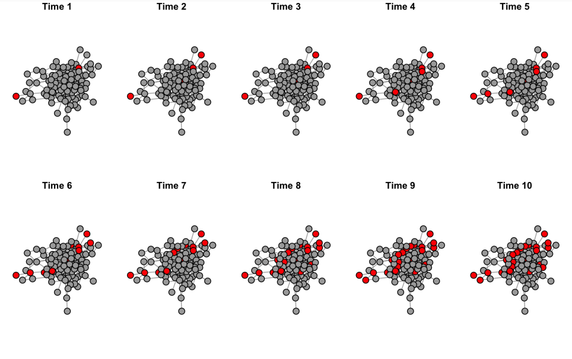


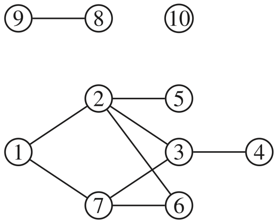
Figure reference: https://dshizuka.github.io/

Basics of network theory

- **Communication centrality**: At each time period, a probability of passing information from an individual node to its adjacent neighbors can vary by a subject's outcome.
- **Diffusion centrality**: At each time period, a *homogeneous* probability p is assumed, regardless of a subject's characteristics and outcomes.
- The diffusion centrality of node u with a probability of information passing p and the process time period T is given by

$$\sum_{j=1}^{N} \sum_{t=1}^{T} (p\mathbf{E})_{uj}^{t}.$$

Basics of network theory



Node	Degree	Betweenness	Diffusion*
1	2	0.67	1.23
2	4	7.50	1.92
3	3	5.67	1.62
4	1	0.00	0.57
5	1	0.00	0.66
6	2	0.67	1.23
7	3	2.50	1.53
8	1	0.00	0.39
9	1	0.00	0.39
10	0	0.00	0.00

Figure 2. A hypothetical network with $N = 10$ nodes. The table shows the centrality measures using degree, betweenness, and diffusion centralities ($p = 0.3$, $T = 2$).

Table of Contents

Motivating example: TRIP

As a motivating example, we discuss the Transmission Reduction Intervention Project (TRIP), which was conducted in Athens, Greece, between 2013 and 2015.

- The goal of this project is to examine the impact of behavioral interventions on HIV-related outcomes within HIV-risk networks of people who inject drugs (PWID).
- Eligible participants: Grown-ups (18 y.o. or older) who were willing to answer a questionnaire.
- Nodes: Participants and their contacts, included up to two waves of recruitment.
- Edges: Sharing of injection equipment or having unprotected sexual intercourse together.

Motivating example: TRIP

The TRIP network consists of $N = 277$ participants with 542 undirected ties, excluding all singleton nodes.

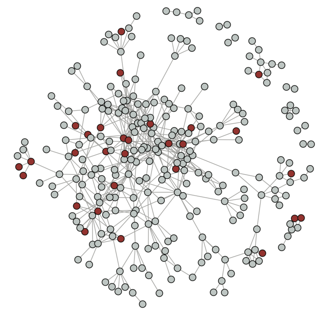


Figure 1. TRIP network with red nodes representing the participants who received the community alert.

Motivating example: TRIP

To reduce HIV-risk behaviors, the researchers distributed community alerts to a subset of participants.

- **Intervention (Treatment)**: Receiving the community alert.
  - A total of 29 (10.5%) of participants received these alerts.
- **Outcome**: Indicator of HIV-risk behavior (sharing injection equipment) at the 6-month follow-up visit.

Question. On which subset of subjects should we intervene to maximize the *collective outcome*?

- We aim to find the most *influential* subset of subjects.
- Are central nodes always the influential ones? Perhaps not.

Table of Contents

Notation and settings

In practice, a single realization of $(\mathbf{Y}, \mathbf{A})$ or $(\mathbf{Y}, \mathbf{A}, \mathbf{Z})$ is observed.

- $\mathbf{A} = (A_1, A_2, \ldots, A_N)^T$: Intervention vector
  - We assume the intervention to be binary, i.e. $\Omega_A = \{0, 1\}$ and $\Omega_{A^N} = \{0, 1\}^N$.
- $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$: Outcome vector
- $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)^T$: Pre-treatment covariates vector

Collective outcomes are any function of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$.

- The simplest example is the average of the outcomes over all nodes.
- Denote this average by $\overline{Y}_{1:N}$.

Notation and settings

Note that the 'no-interference assumption' is violated in network data. Thus, we define potential outcomes for a given intervention assignment vector, $\mathbf{a} = (a_1, a_2, \ldots, a_N)^T$.

- $\mathbf{a}_H := \{\mathbf{a} \in \Omega_{A^N} : a_j = 1 \text{ if } j \in H, a_j = 0 \text{ if } j \notin H\}$
- $Y_i(\mathbf{a})$: Potential outcome of i if subjects were assigned to $\mathbf{a} \in \Omega_{A^N}$
- $\overline{Y}_{1:N}(\mathbf{a})$: Average of potential outcomes over all N subjects in the network under assignment $\mathbf{a}$

A causal measure of influence

Assume, without loss of generality, that we aim to *maximize* the expected average of potential outcomes $\mathbb{E}[\overline{Y}_{1:N}(\mathbf{a})]$. Then, the **problem of finding the most influential subject** is equivalent to the **problem of finding the intervention assignment $\mathbf{a} \in \Omega_{A^N}$ such that $\mathbb{E}[\overline{Y}_{1:N}(\mathbf{a})]$ is maximized**.

- For causal influence of a single node, we consider the assignments in $\Omega_{A_{N,1}} = \{\mathbf{a} \in \Omega_{A^N} : \sum_{i=1}^{N} a_i = 1\}$.
- Likewise, in order to find m influential subjects, consider $\Omega_{A_{N,m}} = \{\mathbf{a} \in \Omega_{A^N} : \sum_{i=1}^{N} a_i = m\}$.

A causal measure of influence

The following definition is now natural:

Definition (Causal influence of nodes)

For a set of nodes $H \subseteq V(\mathcal{G})$, the causal influence of nodes in H can be defined as

$$\tau(H) = \mathbb{E}\left[\overline{Y}_{1:N}(\mathbf{a}_H)\right] = \mathbb{E}\left[\overline{Y}_{1:N}(a_{j,j\in H} = 1, a_{j,j\notin H} = 0)\right] \qquad (1)$$

where the expectation is taken over all N nodes in the network.

Note that such a notion of causal influence remains the same in any network data, regardless of network structures and diffusion processes.

Congruence with the centrality measures

How does our measure of causal influence relate to existing centrality measures? We explore the theoretical conditions under which several well-known centrality measures can correctly identify the causal influence of a single node, namely *out-degree centrality*, *betweenness centrality*, and *diffusion centrality*.

Congruence with the centrality measures

We express the potential outcome of each node i through the following structural causal model:

$$Y_i(\mathbf{a}_{\{k\}}) = \delta_i + \alpha_i \mathbb{1}(a_i = a_k) + \sum_{j \neq i} \beta_{ji} \mathbb{1}(a_j = a_k) + \epsilon_i,\ k \in \{1, \ldots, N\},$$

- $\delta_i = Y_i(\mathbf{0}_N)$: baseline outcome of node i
- $\alpha_i = Y_i(\mathbf{a}_{\{i\}}) - Y_i(\mathbf{0}_N)$: direct effect of intervention on itself,
- $\beta_{ji} = Y_i(\mathbf{a}_{\{j\}}) - Y_i(\mathbf{0}_N)$: node j's effect on node i's outcome,
- ϵ_i are mean-zero random errors.

Assumptions: Outcome variable is continuous, and a higher value of $\mathbb{E}[\overline{Y}_{1:N}(\mathbf{a})]$ implies higher influence. The adjacency matrix $\mathbf{E}$ is given and not affected by interventions.

Congruence with the centrality measures

Proposition 1. (Out-degree centrality)

If α_i is homogeneous across all i and that $\beta_{ji} = 0$ when $e_{ji} = 0$ and $\beta_{ji} = \beta > 0$ when $e_{ji} = 1$, then higher out-degree centrality implies higher influence of τ, and vice versa.

Proposition 2. (Betweenness centrality)

If α_i is homogeneous across all i and that $\beta_{ji} = \beta \sum_{k \neq i,j} \phi_{ki}(j)/\phi_{ki}$ with $\beta > 0$, then higher betweenness centrality implies higher influence of τ, and vice versa.

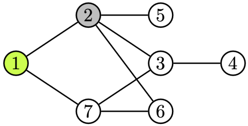
Proposition 3. (Diffusion centrality)

If α_i is homogeneous across all i and that $\beta_{ji} = \beta \{\sum_{t=1}^{T} (p\mathbf{E})^t\}_{ji}$, then higher diffusion centrality at given p and T implies higher influence of τ, and vice versa.

Congruence with the centrality measures

Proposition 2. (Betweenness centrality)

If α_i is homogeneous across all i and that $\beta_{ji} = \beta \sum_{k \neq i,j} \phi_{ki}(j)/\phi_{ki}$ with $\beta > 0$, then higher betweenness centrality implies higher influence of τ, and vice versa.



i	j	k	ϕ_{ki}	$\phi_{ki}(j)$	$\phi_{ki}(j)/\phi_{ki}$
1	2	3	2	1	0.5
		4	2	1	0.5
		5	1	1	1
		6	2	1	0.5
		7	1	0	0

Figure S1. When $i = 1$ and $j = 2$, $\sum_{k \neq i,j} \phi_{ki}(j)/\phi_{ki} = 2.5$, and this determines the influence of node $j = 2$ on node $i = 1$.

Congruence with the centrality measures

The assumptions in the propositions may not be realistic in some settings. In the TRIP study context,

- All of the propositions imply that the direct effect of the community alert is the same for all participants.
- Proposition 1 assumes that the community alert received by participant j only affects its adjacent node i, and such an adjacent effect is homogeneous across all adjacent pairs.
- Proposition 2 assumes that the alert received by participant j affects participant i's outcome even if they are not adjacent, as long as j is on the shortest path connecting i and other nodes.
- Proposition 3 assumes that the intervention effect can spread up to T geodesic distances with the probability of passing the intervention effect being homogeneous.

Table of Contents

Simulation

In order to investigate how well centrality measures agree with the causal influence measure of τ, we consider three data-generating models, which correspond to the assumptions in Propositions 1-3.

(i) Homogeneous direct interference:

$$Y_i(\mathbf{a}) = \delta_i + \alpha \mathbb{1}(a_i = 1) + \beta \sum_{j \neq i} e_{ji} \mathbb{1}(a_j = 1) + \epsilon_i$$

(ii) Traffic-dependent process:

$$Y_i(\mathbf{a}) = \delta_i + \alpha \mathbb{1}(a_i = 1) + \beta \sum_{j \neq i} \{ \sum_{k \neq i,j} \phi_{ki}(j)/\phi_{ki} \} \mathbb{1}(a_j = 1) + \epsilon_i$$

(iii) Homogeneous diffusion process:

$$Y_i(\mathbf{a}) = \delta_i + \alpha \mathbb{1}(a_i = 1) + \beta \sum_{j=1}^{N} \{ \sum_{t=1}^{T} (p\mathbb{E})^t \}_{ji} \mathbb{1}(a_j = 1) + \epsilon_i$$

Simulation

Simulation settings:

- We generate the three different diffusion models upon the network with $N = 277$ non-isolated subjects, replicating each model $r = 500$ times.
- For each replicate, we randomly perturb the network structure, adding and removing 20% random ties while keeping the node size at $N = 277$.
- Along with the three centralities, we also calculate a random selection of nodes as the influential ones.
- To remove the impact of random errors, we set $\epsilon_i = 0$ in our primary simulation.

Simulation results

The simulation results are illustrated in terms of Spearman's rank correlation and node ranking.

	Degree	**Betweenness**	**Diffusion**	**Random**
Process (i)	**1.00** (0.00)	0.71 (0.04)	0.94 (0.01)	0.01 (0.06)
Process (ii)	0.71 (0.04)	**1.00** (0.00)	0.52 (0.03)	0.00 (0.06)
Process (iii)	0.94 (0.01)	0.53 (0.03)	**1.00** (0.00)	0.01 (0.06)

Table 2. Average of the Spearman's rank correlation (ρ) and its standard error between the causal influence τ and each centrality measure. The homogeneous diffusion process parameters are $p = 0.3$, $T = 5$.

Simulation results

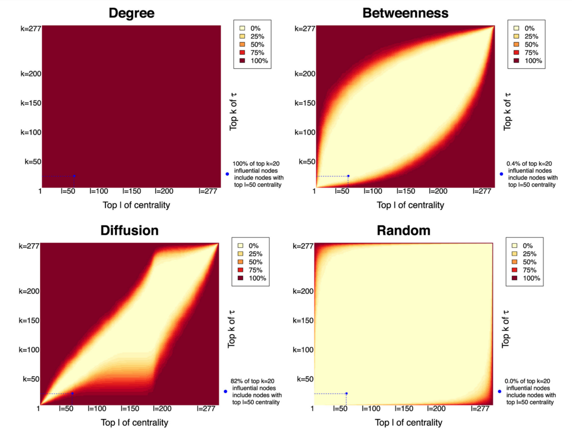


Figure 3. Simulation results under (i) homogeneous direct interference. Each matrix contains 277×277 cells. Each cell illustrates how often the top l of influential nodes established through each centrality metric are completely contained in the top k causally influential nodes for $l \geq k$.

Simulation results

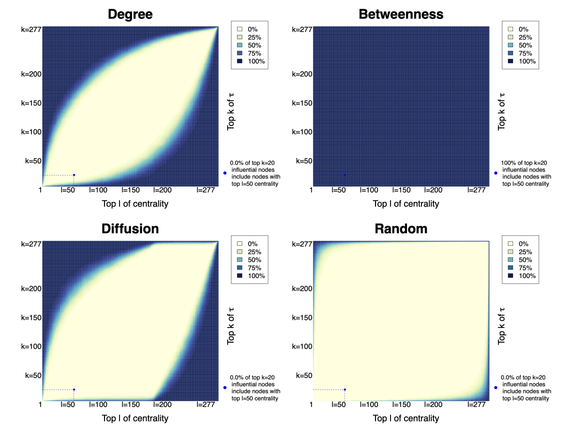


Figure S2. Simulation results under the (ii) traffic-dependent process.

Simulation results

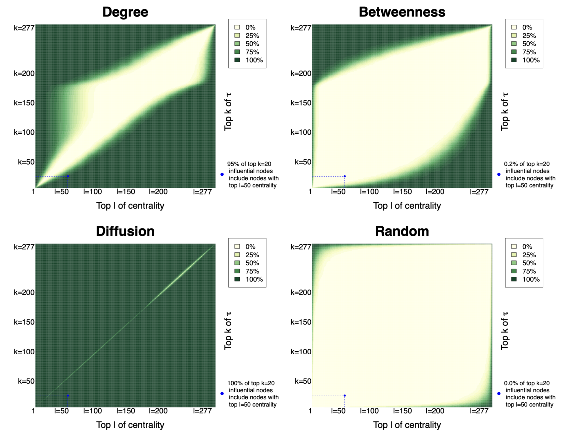


Figure S3. Simulation results under the (iii) homogeneous diffusion process ($p = 0.3$, $T = 5$).

Simulation results

Summary of main results:

- The results suggest that the three centrality metrics are likely to fail to capture the causal measure of influence τ except for one particular data-generating process for each centrality.
- Even so, each of the centrality metrics is better than random selection.
- Therefore, to have each centrality as a valid measure of influence, we may require stringent assumptions on the causal mechanism underlying how the intervention impacts the collective outcome.

Table of Contents

Example analysis: Finding influential participants in TRIP

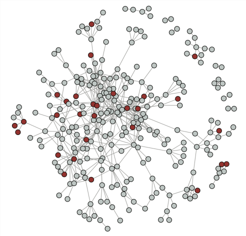
We use three centrality measures (out-degree, betweenness, and diffusion) to analyze the TRIP network, each of which can provide a valid influence measure of a single node under specific causal assumptions in Propositions 1–3.

	Degree	**Betweenness**	**Diffusion ($T = 5$)**
Degree	1.00	0.72	0.86
Betweenness	.	1.00	0.53
Diffusion ($T = 5$)	.	.	1.00

Table 4. The Spearman's rank correlation between the centralities in the TRIP network.

Example analysis: Finding influential participants in TRIP

Six months after the intervention, the number of participants engaging in HIV-risk behavior decreased from 207 (74.7%) to 94 (42.5%).



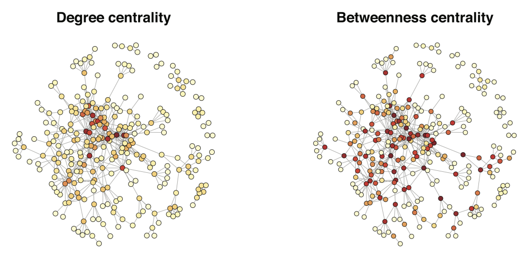


Figure 1. (Left) Revisited.

Figure 4. (Right) The degree and betweenness centralities applied to the TRIP network. The nodes are shaded based on the 20-quantiles of each centrality metric, and a darker shading indicates higher centrality.

Example analysis: Finding influential participants in TRIP

- To guarantee that each of the three centralities are valid, the homogeneous direct effect (α_i) assumption is required.
- Given the 6-month time difference between intervention assignment and evaluation of the targeted outcomes, the *homogeneous direct interference* assumption may not be well supported in this study. Instead, the *traffic-dependent process* may be more reasonable.
- If researchers know the maximum geodesic distance over which the intervention effect might propagate, the *diffusion centrality* measure can be a better alternative.
- Even if the conditions required by each diffusion process are satisfied, the centralities may not accurately capture causal influence, as interventions can alter network structures.

Table of Contents

Main findings of the paper

- In this paper, we defined the influence of subjects on a network as a causal effect on collective outcomes using a potential outcomes framework, thereby making a clear distiction between *importance* and *influence*.
- We explored the conditions under which centrality measures coincide with causal influence, and evaluated their effectiveness in capturing influence through a simulation study.
- Most centrality measures used to assess influence are making highly restrictive assumptions about the diffusion mechanism of intervention effects.

Limitations and future works

- Causal effects and measures on network data still poses identification and estimation challenges. The estimation of effects typically relies on parametric assumptions, or requires multiple realizations of $(\mathbf{Y}, \mathbf{A}, \mathbf{Z})$.
- To avoid potential sources of bias that arise in estimation procedures, the authors aim to find randomization designs to identify the influence of each subject in a network.
- Future researchers may also consider using more flexible centrality measures such as random-walk betweenness centrality.