RePBubLik: Reducing the Polarized Bubble Radius with Link Insertions

Shahrzad Haddadan
Dept. of Computer Science
& Data Science Initiative
Brown University
Providence, RI, USA
shahrzad_haddadan@brown.edu

Cristina Menghini
DIAG
Sapienza University
Rome, Italy
menghini@diag.uniroma1.it

Matteo Riondato
Dept. of Computer Science
Amherst College
Amherst, MA, USA
mriondato@amherst.edu

Eli Upfal
Dept. of Computer Science
Brown University
Providence, RI, USA
eli@cs.brown.edu

Democracy begins in conversation — John Dewey (attr.)

ABSTRACT

The topology of the hyperlink graph among pages expressing different opinions may influence the exposure of readers to diverse content. Structural bias may trap a reader in a "polarized" bubble with no access to other opinions. We model readers' behavior as random walks. A node is in a "polarized" bubble if the expected length of a random walk from it to a page of different opinion is large. The structural bias of a graph is the sum of the radii of highly-polarized bubbles. We study the problem of decreasing the structural bias through edge insertions. "Healing" all nodes with high polarized bubble radius is hard to approximate within a logarithmic factor, so we focus on finding the best k edges to insert to maximally reduce the structural bias. We present RePBubLik, an algorithm that leverages a variant of the random walk closeness centrality to select the edges to insert. RePBubLik obtains, under mild conditions, a constant-factor approximation. It reduces the structural bias faster than existing edge-recommendation methods, including some designed to reduce the polarization of a graph.

CCS CONCEPTS

• **Theory of computation** → *Random walks and Markov chains*; **Graph algorithms analysis**; • **Information systems** → *Social networks*; **Social recommendation**.

KEYWORDS

Bias, Fairness, Polarization

ACM Reference Format:
Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. 2021. RePBubLik: Reducing the Polarized Bubble Radius with Link Insertions. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441825

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441825

1 INTRODUCTION

The World Wide Web often contains thousands or even millions of pages on every topic, covering the whole spectrum of opinions. Exposure to *diverse content* is necessary to obtain a complete picture about a topic. This exposure depends on the hyperlinks connecting the pages to each other. It can be argued that enabling easier access to diverse content improves society as it creates a more informed and less polarized general public [11]. Indeed politicians have strongly promoted and even requested that audiences are exposed to varied content [36].

The fact that diverse information is easily *available* does not imply that *exploring* such diverse information is easy. Rather, echo chambers and polarization on (social) media and blogs [1, 17, 25] keep the user in a *homogeneous bubble*, exposing them only to agreeable information [8], and leading to conflicts between users in different bubbles [18, 35].

A web user can freely click on any hyperlink on the page they are currently visiting, but the choice of which hyperlinks to include in the page is with the website owner or editor, who, if not careful, may stop the user from being exposed to diverse opinions. In other words, the hyperlink topology of a website may suffer from *structural bias* that traps the user in a bubble of one-sided content without them knowing [45, 56]. For example, structural bias on topic-induced networks, such as Wikipedia topic-induced subgraphs, prevents users from building a well-rounded knowledge about the topic. On query-/user-induced recommendation networks such as those on Amazon and YouTube, structural bias hinders the discovery of diversified content, reducing serendipity [3, 30, 34]. Structural bias thus limits the user's freedom while navigating the Web.

Socially-minded website editors would try to *minimize such structural bias* by adding appropriate links to pages in the website. As an editor can only modify a few pages and add only a few links to each edited page, they need to carefully choose what page p to edit, and what pages to link to from p. The goal of our work is to develop an algorithm that can give editors recommendations for what links to add in order to reduce the structural bias. There are key technical challenges that must be solved in order to give effective link recommendations: *1.* quantify the structural bias of a page p, i.e., how hard it is to reach pages of a different opinion from p; and *2.* decide which of these pages should be linked from p. Existing approaches to link recommendation, such as those based on vertex similarity, fall short at this task because they are oblivious to the network structural bias. Their recommendations often *increase* the

bias, rather than decreasing it [49], especially for highly controversial topics such as political blogs (see also our experimental results in Sect. 6). Thus, there is a need for a radically different approach to suggest links that decrease the structural bias.

Contributions. We study the problem of reducing the structural bias of a graph by adding edges, and propose an algorithm, RePBubLik, to suggest such edges. Our contributions are the following.

- We consider directed graphs with vertices of two colors, representing a network of webpages on the same topic, with the two colors identifying the two opposite opinions on the topic, and edges representing links between pages. We define the *(Polarized) Bubble Radius* (BR) of a vertex p as a novel measure to quantify the structural bias of p (see Def. 4.1), based on a task-specific variant of the hitting time for random walks, which models the navigation of a user on the web [23, 24]. The BR is the expected number of steps to go from p to a page of different opinion, and can be easily estimated with a sampling-based approach with probabilistic guarantees (Lemma 4.3), which enables us to tackle the first of the key challenges.
- We define the *structural bias* of a graph G as the sum of the BRs of vertices with high BR (Eq. 1). Completely removing the bias is APX-hard by reduction from set cover (see Lemma 4.5). We therefore state the *k-edge structural bias decrease* problem as the task of finding the set of k pairs of vertices of different color such that adding the edge between the vertices in each pair would *maximally* decrease the structural bias, over all possible sets of k pairs (see Prob. 2 and Thm. 5.1). This problem connects two areas: link recommendation and polarization reduction.
- We present RePBubLik, an efficient approximation algorithm for the k-edge structural bias decrease problem, that recommends the addition of k edges between vertices of different color. Under mild conditions, the resulting decrease of the structural bias is within a constant factor of the optimal. Website editors have limited control on the probability that a newly added edge will be traversed by the users, so our algorithm makes no assumption or impose any restriction on it, as this probability is essentially external. At the core of RePBubLik is an analysis of the submodularity of the objective function (see Lemma 5.6), combined with the use of a task-specific variant of random-walk closeness [63], a well-established centrality measure. RePBubLik requires good estimations of the random walk closeness, so we also give an approximation algorithm for this quantity (see Lemma 3.1).
- We evaluate RePBubLik on eight real datasets. We compare it to baselines and existing methods for edge recommendation either designed with the goal of reducing the controversy of a graphs [26] or with the more general purpose of completing the network's link structure [31]. Our algorithm leads to a faster reduction of the average BR (i.e., requiring fewer edge insertions) than existing contributions.

Due to space limitations, many of our proofs are in the appendix of the extended online version [32].

2 RELATED WORK

Polarization has long been studied in political science [33, 60], and the recent diffusion of (micro-) blog and social media platforms brought the issue to the attention of the broad computer science community. Many works focused on showing the existence of polarization on these platforms [1, 17, 18, 25, 46], and on modeling, quantifying, and reducing polarization [2, 7, 9, 16, 26–29, 38, 39, 41, 41, 42, 48–50], or the glass ceiling effect [57–59]. The literature is rich, to the point that times seems ripe for an in-depth survey on the topic. Due to space limitations, we discuss here only the relationship between our work and the most relevant algorithmic contributions to polarization reduction [7, 9, 16, 26, 28, 42, 44, 45, 49, 59].

A first important difference of our work with respect to most previous contributions is that they consider a network of *users*, with edges representing notions such as friendship or endorsement (e.g., retweets) [7, 9, 16, 26, 28, 42, 49, 59]. We focus instead on networks of *content*, such as web pages linked to each other, or products that are connected when similar. This deep difference makes our contribution quite orthogonal to the ones in these previous works: we focus on the polarization that is introduced by the topology of the network, rather than on the polarizing effect of content on users or on the effect of users on each other. We believe both aspects are important, but the structural bias we focus on has only been subject to few studies [44, 45]. These works, relying on the notion of weighted reciprocity, propose a static and dynamic analysis of structural bias on Wikipedia. The measure of structural bias we use is not tailored to a specific website.

A second relevant difference from many previous works is that we consider the "opinion" of a page (i.e., a vertex) to be fixed, as it depends on its content, while many previous contributions consider different models of user opinion dynamics [20, 47] to study the evolution of such opinions as the users are exposed to different content or recommended different friendships. The problem of recommending changes to the content of a page to modify the opinion expressed in it is interesting but outside the scope of our work. Instead, we focus on recommending the addition of links between pages, to reduce the structural bias.

An interesting line of work studies how to reduce polarization in the content seen by the users, by adapting information diffusion approaches through better selection of the seed set for cascades [7, 9, 28, 42, 58], or by directly acting on recommendation systems [55]. These methods can not be adapted to the problem we study, as they do not act on the graph of content, but on that of users.

The most similar methods to ours are those that act on the structure of the graph [16, 26, 49, 59], although as we mentioned, they consider a network of users, not of content. Musco et al. [49] propose a network-design approach: they aim to find the best set of edges between vertices such that the resulting graph would minimize both disagreement and polarization. Rather than a "design-from-scratch" approach, which seems mostly of theoretical relevance, we consider instead a practical incremental approach that suggests modifications to an existing network. Like us, Garimella et al. [26] consider a graph polarization measure based on random walks [29]. This measure essentially quantifies the probability that a user of one opinion is exposed to content from a user of a different opinion, thanks to a chain of retweets (represented by the

random walks). The measure is based on a variant of personalized PageRank for sets of users with different opinions. The task requires to recommend new edges, i.e., retweets, to increase this probability. Our measure of structural bias is instead defined on the basis of the (Polarized) Bubble Radius (BR) (Def. 4.1), which is a vertex-dependent measure that represents the expected number of steps, for a user starting at the page represented by vertex v, to reach, with a random walk, a vertex with color different from v, representing a page expressing a different opinion. Our measure is appropriate for our task of suggesting new edges to make it easier for user to reach pages of different opinions. In Sect. 6 we compare our approach to that of Garimella et al. [26].

An important line of work in graph analysis and mining looked at manipulating the topology to modify different interesting characteristic quantities of the graph, such as shortest paths and related measures [22, 51, 52, 54], various forms of centrality [4, 13, 19, 40, 43, 53, 62], and more [5, 6, 14, 61, 64]. Despite the fact that we consider a specific centrality to choose the source of the added edges, these methods cannot be used to solve our task of interest.

Another body of work related to ours are those which estimate graph properties using random walks [10, 12, 15, 21]. The studied properties are not defined based on random walks, rather random walks are used as a tool to estimate them. Here based on random walks, we define a new property for networks: *the structural bias*, and we use random walks to estimate it.

3 PRELIMINARIES

Let $G = (V, E)$ be a directed weighted graph with $|V| = n$ vertices, such that no vertex $v \in V$ has only incoming edges and no outgoing edges. V is partitioned in two disjoint sets R and B (i.e., $R \cap B = \emptyset$ and $R \cup B = V$), called "red" or "blue" vertices, respectively. We denote the color of a vertex v by $\mathrm{c}(v)$ and its opposite color by $\bar{\mathrm{c}}(v)$. The sets of all *other* vertices of the same color as v is denoted as C_v and the sets of all vertices of color different than v is denoted as $\bar{C}_v$.

The edge weights are *transition probabilities*, as follows. Let M be a $n \times n$ right-stochastic *transition matrix* associated to G, i.e., a matrix such that each entry $m_{i,j}$ is a probability, with $m_{i,j} = 0$ if $(i, j) \notin E$, and such that $\sum_{j=1}^{n} m_{i,j} = 1$.

We are interested in random walks on the graph G using the transition matrix M. Intuitively, a random walk starting at a vertex v explores the graph by choosing at each step an outgoing edge from the current vertex, with probability equal to the weight of such edge, independently from previous choices. Let $S \subseteq V$ and $v \in V$. Let $T_v(S)$ be the random variable indicating the first instant when a random walk from v hits (i.e., reaches) any vertex in S. The quantity $\mathbb{E}_G[T_v(S)]$ is known as the *hitting time of S from v*, where the expectation is over the space of all random walks on G starting from v, with transition probabilities given by M.

Variants of random walks, such as random walks with restarts or with back button, are widespread models for network exploration [23, 24]. It is realistic to assume that there is an upper bound t, which we call the *exploration factor*, on the length of a walk performed by the users. For example, we can assume that there is an upper limit on the number of pages that a user will visit one after the other in a browsing session. The value of the parameter t can be derived, for example, from traces of visits. In most practical cases, t is likely to be bounded by a polylogarithmic quantity in the number of nodes, if not a constant.

For a random walk starting from $v \in V$, given a set $S \subseteq V$, we define the random variable $T_v^t(S)$ as $\min\{t, T_v(S)\}$. This variable is more appropriate for measuring the length of browsing sessions, which have bounded length, than the unbounded length classically used when discussing random walks.

For a graph Z, any vertex u, and any set S of vertices, let $u \overset{\text{cond}}{\underset{Z}{\rightsquigarrow}} S$, denote the event that a random walk in Z from u hits a vertex in S without first visiting any vertex in $\bar{C}_u$ and while satisfying the condition cond on the number of steps needed to hit S. For example, $u \overset{<t}{\underset{Z}{\rightsquigarrow}} S$ is the event that a random walk in Z from u hits a vertex in S in *less than t steps*, without first visiting any vertex in $\bar{C}_u$. We denote the complementary event as $u \overset{\text{cond}}{\underset{Z}{\not\rightsquigarrow}} S$.

3.1 Random-Walk Closeness Centrality

We adapt the definition of the standard random-walk closeness centrality [63] to bounded random walks so that the contribution to the centrality of v by vertices that do not reach v in less than t' steps (in expectation) is zero, for any t'.

Random-walk closeness centrality (bounded form). For a vertex $v \in V$, and any t', the *t'-bounded* Random Walk Closeness Centrality (RWCC) measure with respect to subset $S \subseteq V$ is

$$\mathrm{r}^{t'}(v; S) \doteq \frac{1}{|S|} \sum_{w \in S} \left(t' - \mathbb{E}_G\left[T_w^{t'}(v)\right] \right)$$
$$= \frac{1}{|S|} \sum_{w \in S} \sum_{i=1}^{t'} (t' - i) \mathbb{P}\left(w \overset{=i}{\underset{G}{\rightsquigarrow}} v \right) \quad .$$

Computing the exact RWCC is expensive. To estimate $\mathrm{r}^{t'}(v; S)$, we pick z vertices $\{w_i\}_{i=1}^{r}$ u.a.r. from S, and run some κ random walks to obtain an estimate $\bar{h}_{w_i}$ of $\mathbb{E}_G\left[T_{w_i}^{t'}(v)\right]$ for each w_i. The quantity $\bar{r}(v) \doteq t' - 1/z \sum_{i=1}^{z} \bar{h}_{w_i}$ is a good approximation of $\mathrm{r}^{t'}(v; S)$.

Lemma 3.1. *Let $z \geq (t'/2\varepsilon)^2 \delta^{-1}$. Then*

$$\mathbb{P}\left(|\bar{r}(v) - \mathrm{r}^{t'}(v; S)| \geq \varepsilon \right) \leq \delta \quad .$$

4 BUBBLE RADIUS AND STRUCTURAL BIAS

We introduce the *(Polarized) Bubble Radius* to quantify how *likely* users starting their random walk on a vertex $v \in V$ of one color, are to hit a vertex of the other color in at most t steps.

Definition 4.1. The *(Polarized) Bubble Radius (BR)* $\mathrm{B}_G^t(v)$ *of v with exploration parameter t* is

$$\mathrm{B}_G^t(v) \doteq \mathbb{E}_G\left[T_v^t\left(\bar{C}_v\right)\right] \quad .$$

A random walk starting at a vertex v with high BR is unlikely to hit a vertex in $\bar{C}_v$ in fewer-than-or-exactly t steps. The following lemma formalizes this idea on common models for web browsing (random walks with restarts or with back button [23, 24]).

Lemma 4.2. *Let $r \in \mathbb{N}$, and consider a user who starts their random walk at $v \in V$ and may either restart their walk from v or hit the*

back button up to r times. Let $\mathcal{T}_v$ be the random variable denoting the number of steps such user takes to hit a vertex in $\bar{C}_v$. If $\mathrm{B}^t_G(v) \geq t(1 - 1/8r)$, then $\mathbb{P}(\mathcal{T}_v \leq t/2) \leq 1/4$. If instead $\mathrm{B}^t_G(v) \leq b$ for some $b > 0$, then $\mathbb{P}(\mathcal{T}_v > 4br) \leq 1/4$.

Given t, it is easy to estimate $\mathrm{B}^t_G(v)$ for each vertex $v \in V$ by sampling random walks from v. The following result, whose proof uses the Hoeffding's bound and the union bound, shows the trade-off between the number of sampled random walks and the accuracy in estimating the BR of v.

Lemma 4.3. *For each $v \in V$, let $w_1^{(v)}, w_2^{(v)}, \ldots, w_r^{(v)}$ be r random walks from v and stopped either when they hit a vertex of color $\bar{c}(v)$ or when they run for t steps, whichever happens first. For $i = 1, \ldots, r$, let $b_i^{(v)}$ be the length of random walk $w_i^{(v)}$. Let*

$$\bar{B}(v) \doteq \frac{1}{r} \sum_{i=1}^{r} b_i^{(v)} \quad .$$

Let $\varepsilon, \delta \in (0, 1)$. If $r \geq \frac{t^2}{\varepsilon^2} \ln \frac{2n}{\delta}$, then

$$\mathbb{P}\left(\exists v \in V \text{ s.t. } |\mathrm{B}^t_G(v) - \bar{B}(v)| > \varepsilon\right) < \delta,$$

where the probability is over the choice of the random walks.

In the rest of the work, we assume for simplicity to have access to the *exact* BR of every vertex. The above result makes this assumption reasonable because computing approximations of extremely high quality is relatively inexpensive.

The structural bias. On the basis of the BR, we define two sets of vertices: *cosmopolitan* and *parochial*. Given two reals b and r with $1 \leq b < r \leq t$, the set $\mathcal{Z}(G)$ of *cosmopolitan* vertices contains all and only the vertices in G with BR *at most* b, and the set $\mathcal{P}(G)$ of *parochial* vertices contains all and only the vertices in G with BR at least r. For ease of notation, we do not include b and r in the notation for $\mathcal{Z}(G)$ and $\mathcal{P}(G)$. In the rest of this work, we assume for simplicity $b = 2$ and $r = t/2$, but this assumption can be easily removed. $\mathcal{Z}(G)$ and $\mathcal{P}(G)$ are *disjoint*, but they do not necessarily form a partitioning of V. We will often consider the partitioning of $\mathcal{P}(G)$ by color, i.e., the two sets $\mathcal{P}_R(G)$ and $\mathcal{P}_B(G)$, containing the parochial vertices of color R or B respectively.

Definition 4.4. The *structural bias* $\rho(G)$ of G is the sum of the BRs of the parochial nodes of G, i.e.,

$$\rho(G) \doteq \sum_{v \in \mathcal{P}(G)} \mathrm{B}^t_G(v) \quad . \tag{1}$$

It is reasonable to consider only the parochial nodes in the definition of structural bias because they are the ones such that a random walk from them is very unlikely to hit any vertex of color different than the starting vertex (see also Lemma 4.2).

Our goal in this work is to find a set of edges with extrema of different color whose addition to G would decrease the structural bias of the network. It is reasonable to only consider edge with extrema of different color, as they are always preferable (i.e., will result in a higher decrease of the structural bias) than edges with monochromatic extrema: the addition of the new edge can only have positive impact on the parochial vertices of the same color as the source, and has no impact on the parochial vertices of the other color. If we could add *any number* of such edges to G, it would be easy to bring the structural bias of G to zero, as there would be no parochial nodes left. This assumption is not realistic: the number of links that a website editor can add to a single page and to the whole graph is limited by many factors, such as the fact that a human-readable page cannot have too many links, and the fact that the editor can only spend a limited time on this activity. Nevertheless, ideally one would want to solve the following problem.

Problem 1. Given a color $C \in \{R, B\}$, find the smallest set A of pairs of distinct edges $(v, w) \notin E$ with $\mathrm{c}(v) = C$ and $\mathrm{c}(w) \neq C$ such that, for the graph $G_{\text{new}} = (V, E \cup A)$ it holds $\mathcal{P}_C(G_{\text{new}}) = \emptyset$.

Lemma 4.5. *Problem 1 is NP-hard and APX-hard.*

5 REDUCING THE BR WITH INSERTIONS

Since Prob. 1 is hard to even approximate (Lemma 4.5), we seek to answer a close relative (Prob. 2). We first introduce a set of measures to capture the change in the BRs of the (original) parochial nodes of G after edge insertions. Let G_{new} be obtained from G by inserting a set Σ of directed edges between nodes of different colors, with each inserted edge $e = (v, w)$ having weight m_e (also denoted as m_{vw}). For a set U of vertices, we define the *gain* of U due to Σ as

$$\Delta(G, U, \Sigma, \{m_e\}_{e \in \Sigma}, t') \doteq \frac{1}{|U|} \sum_{u \in U} \left(\mathrm{B}^{t'}_G(u) - \mathrm{B}^{t'}_{G_{\text{new}}}(u)\right) \quad .$$

When adding an edge to the graph, we also have to decide its weight. It seems excessive to assume complete freedom in choosing the weight. We make the assumption that the weight m_{vw} of an edge (v, w) that we would like to add is given to us by an oracle which computes m_{vw} only as a function of v and of information *local* to v (e.g., its out-degree) obtained from G and potentially a set of other edges (and their weights) that we want to add from v. The weight m_{vw} is the probability that a random walk arriving at v will move to w in the next step. When adding (v, w) with weight m_{vw}, the other edges outgoing from v have their weights multiplied by $1 - m_{vw}$ to ensure that the sum of the weights of the edges leaving v is 1.

The problem we want to solve then is the following.

Problem 2. In graph G, let C be either blue or red. Find a set $\Sigma = \{(v_i, w_i)\}_{i=1}^{k}$ of k edges whose source vertices all have color C and all destination vertices have the other color, that maximizes $\Delta(G, \mathcal{P}_C(G), \Sigma, \{m_e\}_{e \in \Sigma}, t)$.

RepBubLik (Alg. 1) is our algorithm to approximate Prob. 2. Before describing it in detail, we give an intuition of its workings, and present the theoretical results that guided its design. Specifically, since our objective function is *monotonic and submodular* (Lemma 5.6), we can greedily choose the edges to be added one by one. Due to our oracle assumption on the weights, any vertex of color different than the source can be picked as the target of the added edge, so the problem essentially *reduces to finding the sources for the edges to be added.* Lemma 5.4 quantifies the gain when picking each source according to a specific measure depending on the bounded RWCC and on the oracle-given weight that only depends on the source. In Lemma 5.5 we show that under mild conditions this choice is constantly close to an optimal choice. The following theorem states the approximation qualities of RepBubLik.

Theorem 5.1. *Let Σ be the output of RepBubLik and* OPT *be the optimal solution to Prob. 2. Let $\Delta_\Sigma = \Delta(G, \mathcal{P}_C(G), \Sigma, \{m_e\}_{e\in\Sigma}, t)$*
Then

$$\Delta(G, \mathcal{P}_C(G), \text{OPT}, \{m_e\}_{e\in\text{OPT}}, t) \leq (4\gamma(G)+1)\left(1+\frac{1}{e}\right)\Delta_\Sigma,$$

where $\gamma(G)$ is the maximum over all $u \in V$ of sum of the probabilities, for $i = 0, \ldots, t-1$, that a random walk starting at u visits u at step i without first visiting a vertex in $\bar{C}_u$ (see also (2)), which is a constant for many graphs.

We now proceed towards presenting lemmas which together provide a proof for Thm. 5.1. Lemmas 3.1 and 4.3 provide bounds of order $\Theta(nt^2)$ on the runtime of the pre-processing phases of RepBubLik. Therefore for small values of t, RepBubLik is more efficient than algorithms that compute hitting times using the Laplacian, which need $\Omega(n^3)$ steps.

For any vertex v, and $0 \leq i \leq t$, let $\Psi_v(i)$ be the probability that a random walk (in G) from v visits v at step i before reaching a vertex in $\bar{C}_v$ (it holds $\Psi_v(0) = 1$ and $\Psi_v(1) = 0$ for every v). For any $t' \leq t$, let

$$\mathcal{F}_{t'}(v) = \sum_{i=0}^{t'-1} \Psi_v(i) \quad . \tag{2}$$

The following lemma shows upper and lower bounds to the change in the bubble radius of a vertex when an new edge from it is added to the graph.

Lemma 5.2. *Let $v \in \mathcal{P}(G)$, $w \in \bar{C}_v$ and $t' \leq t$. Let G_{new} be the graph obtained after adding $e = (v, w)$ to G, with weight m_e. The gain $\Delta(G, v, e, m_e, t')$ is such that*

$$\left(\mathrm{B}_G^{t'}(v) - 1\right) m_e \leq \Delta(G, v, e, m_e, t') \leq \mathcal{F}_{t'}(v)\left(\mathrm{B}_G^{t'}(v) - 1\right) m_e \quad .$$

Decreasing the BR of v decreases the BRs of vertices in C_v close to v, and thus the whole network. Lemma 5.3 quantifies this change.

Lemma 5.3. *Let $e = (v, w)$ be the edge with weight m_e added to G to obtain G_{new}. For any other vertex $u \in \mathcal{P}_{C_v}(G)$, it holds*

$$\Delta(G, u, e, m_e, t) = \sum_{i=1}^{t-2}\left(\Delta(G, v, e, m_e, t-i)\mathbb{P}\left(u \overset{=i}{\underset{G}{\rightsquigarrow}} v\right)\right) \quad .$$

Recall that our greedy choice is to identify a node v that maximizes the gain $\Delta(G, \mathcal{P}(G), (v, w), m_v, t)$ where w is any vertex in $\bar{C}_v$. Lemma 5.3 suggests that a good candidate v is a vertex that is likely to be reached by short random walks from many other vertices in $\mathcal{P}_{C_v}(G)$, a property that is captured by the bounded RWCC $\mathrm{r}^{t-2}\left(v; \mathcal{P}_{C_v}(G)\right)$ (Sect. 3.1).

Now, we first quantify the gain for adding an edge from any vertex with RWCC c (Lemma 5.4). Then we show that under mild conditions on the return time of vertices we get a constant approximation by greedily choosing a vertex with maximum RWCC $\times\, m_v$ (Lemma 5.5).

Lemma 5.4. *Let $v \in \mathcal{P}(G)$. Let $w \in \bar{C}_v$, and assume to add the edge $e = (v, w)$ with weight m_e. It holds*

$$\Delta(G, \mathcal{P}_{C_v}(G), e, m_e, t) \geq \frac{m_e}{2}\mathrm{r}^{t-2}\left(v; \mathcal{P}_{C_v}(G)\right) \quad .$$

This lemma suggests that inserting edges from a vertex v with the highest value of $m_v \mathrm{r}^{t-2}(v; \mathcal{P}_C(G))$ *may* result in a larger improvement in the objective function than if we chose a different source. In the next lemma we compare the effect of choosing such sources to the effect of an optimal choice.

Lemma 5.5. *Consider the set $\mathcal{P}_C(G)$ where C is either color. Among all vertices in $\mathcal{P}_C(G)$ let v and* opt *be*

$$\text{opt} = \arg\max_{u\in\mathcal{P}_C(G)} \Delta(G, \mathcal{P}_C(G), e_u, m_u, t),$$
$$v = \arg\max_{u\in\mathcal{P}_C(G)} m_u \mathrm{r}^{t-2}(u; \mathcal{P}_C(G)),$$

where e_u is any potentially inserted edge connecting u to $\bar{C}_u$, and m_u is its weight.[1]
It holds

$$\Delta(G, \mathcal{P}_C(G), e_{\text{opt}}, m_{\text{opt}}, t) \leq (4\gamma(G)+1)\Delta(G, \mathcal{P}(G), e_v, m_v, t),$$

where $\gamma(G) = \max_{u\in G} \mathcal{F}_t(u)$.

If the probability of getting back to u in less than t steps is less than α for some constant α then $\gamma(G) \leq \alpha$. This assumption is realistic since t is usually small and the return time to u is often much larger than t.

Finally, we show that the gain function is monotonic and submodular.

Lemma 5.6. *Let C be either blue or red and $v, u \in \mathcal{P}_C(G)$, and $w_v, w_u \in \bar{C}_v$, such that $e_v = (v, w_v)$ and $e_v = (u, w_u)$ are not existing edges. Let $\Sigma = \{e_v, e_u\}$. It holds*

$$\Delta(G, \mathcal{P}_C(G), e_v, m_{e_v}, t) \leq \Delta(G, \mathcal{P}_C(G), \Sigma, \{m_e\}_{e\in\Sigma}, t),$$

and

$$\Delta(G, \mathcal{P}_C(G), \Sigma, \{m_e\}_{e\in\Sigma}, t) \leq \Delta(G, \mathcal{P}_C(G), e_v, m_{e_v}, t) + \Delta(G, \mathcal{P}_C(G), e_u, m_{e_u}, t) \quad .$$

We are now ready to prove Thm. 5.1.

Proof of Thm. 5.1. Lemma 5.6 shows the monotonicity and submodularity of the objective function. Thus, a greedy algorithm that picks, iteratively, the k best choices over all parochial vertices of color C as the sources of the added edges, will result in a $(1 + 1/e)$-approximation. Lemmas 5.4 and 5.5 show that by choosing a vertex v maximizing $m_v \mathrm{r}^{t-2}(v; \mathcal{P}_C(G))$ among all parochial vertices of color C, we obtain a vertex such that the gain when adding an edge from this source is a $4\gamma(G) + 1$-approximation to the greedy choice. Thus, the correctness of our algorithm is concluded by putting these lemmas together. □

We can now give the details to RepBubLik. The algorithm takes as input the graph G, the number k_C of desired edge insertions, the oracle $\mathcal{W}$ that determines the weights of the new edges, and the set of nodes C. It first creates the empty set Σ_C that will store the edges to be added and then enters a for loop to be repeated for k_C times. At every iteration of the loop, it first computes the BR of every node in C in the graph (denoted in the pseudocode as $G \cup \Sigma_C$) obtained by adding to G the edges currently in Σ_C (with their weights obtained from the oracle $\mathcal{W}_G$) (in practice, the BR is computed using the approximation algorithm outlined in

[1] Our assumption on the oracle giving the weight ensures that m_u only depends on u, not on the target of e_u.

Algorithm 1 RePBubLik

1: **Input**: Graph $G = (V, E)$, desired insertions k_C, oracle $\mathcal{W}_G : V \times 2^{V \times V} \to [0, 1]$, $C \in \{R, B\}$.
2: **Output**: Set Σ_C of k_C edges to be inserted, with their weights.
3: $\Sigma_C \leftarrow \emptyset$
4: **for** $i = 1$ **to** k_C **do**
5: $P \leftarrow$ computeParochials$(G \cup \Sigma_C, C)$
6: $\mathcal{R} \leftarrow$ computeRWCentrality$(P, G \cup \Sigma_C)$
7: $v_i \leftarrow \text{argmax}_{v \in P} \mathcal{R}(v) \times \mathcal{W}_G(v, \Sigma_C)$
8: $u_i \leftarrow$ arbitrary in $\bar{C}_{v_i}$
9: $\Sigma_C \leftarrow \Sigma_C \cup \{(v_i, u_i)\}$
10: **end for**
11: **return** Σ_C

Lemma 4.3). Thanks to this computation, the algorithm obtains (line 5) the set P of parochial nodes in this graph (at the first iteration of the loop $P = \mathcal{P}_C(G)$). It then obtains the centralities values $\mathrm{r}^{t-2}(v; P)$ of every node $v \in P$ (in practice, using the approximation algorithm outlined in Lemma 3.1), storing them in a dictionary $\mathcal{R}$ (line 6). The algorithm then selects the node $v_i \in P$ associated to the maximum quantity $\mathcal{R}(v_i) \times \mathcal{W}_G(v_i, \Sigma_C)$, and arbitrarily picks a node u_i of the opposite color of v_i (i.e., of the color other than C). The directed edge (v_i, u_i) is added to the set Σ_C (lines 7–9). After k_C iterations of the loop, the algorithm returns Σ_C, together with the weights obtained from the oracle.

RePBubLik would require a re-computation of the BRs and of the centralities of all vertices, at every iteration of the loop, which would require to run a very large number of random walks, making it computationally very expensive. We now propose a more practical alternative RePBubLik+, at the price of losing the approximation guarantees. RePBubLik+ only computes $\mathcal{P}_C(G)$ and $\mathcal{R}$ before entering the for loop, and uses the same values throughout its execution, but trades off the consequences of this choice by adding a penalty factor to the objective function involved in the selection of the source vertices for the edges to be added. Specifically, RePBubLik+ chooses v_i (line 7) by maximizing the quantity $\mathcal{R}(v) \times \mathcal{W}_G(v, \Sigma_C)/\eta_v$, where η_v is a penalty factor equals to one plus the number of edges with source v in Σ_C (thus at iteration 1, $\eta_v = 1$ for every node). This penalty factor favours the insertion of edges from nodes that have not yet been altered. Consequently, it indirectly *(1)* handles the possibility that nodes with new edges are no longer parochial, thus we want to avoid to keep adding edges to them; and *(2)* avoids that the new edges are added from a restricted set of nodes, limiting the positive effect of the insertions on $\Delta(G, \mathcal{P}_C(G), \Sigma_C, \{m_e\}_{e \in \Sigma}, t')$.

6 EXPERIMENTAL EVALUATION

The goal of our experimental evaluation is to understand how the addition of the set $\Sigma = \Sigma_R \cup \Sigma_B$ of $K = k_R + k_B$ edges output by RePBubLik+, run separately with $C = R$ and B, affects the structural bias of the network, by computing the gain in the structural bias reduction. In particular, we measure the gain with $\Delta(G, \Sigma)$, introduced in Sect. 5, used here with a simpler notation. We also measure the change $|\mathcal{P}(G)| - |\mathcal{P}(G_{\text{new}})|$ after adding Σ.

Baselines. We compare RePBubLik+ to three different baselines (i.e., simplified variants of RePBubLik+) and to two existing algorithms, described in the following. The first baseline, *PureRandom* (PR) selects the source, and the target, nodes of the new edges uniformly at random from the set $\mathcal{P}_C(G)$ and $\bar{C}$, respectively. The second baseline *Random Top-N Central Nodes (N-RCN)*, given a parameter $N \in (0, 100)$, sorts the nodes in $\mathcal{P}_C(G)$ by descending centrality, and picks, uniformly at random, k_C edges with source in the top-N percent of nodes in $\mathcal{P}_C(G)$. The last baseline, *Random Top-N Weighted Central Nodes (N-RWCN)*, differs from N-RCN as the nodes in $\mathcal{P}_C(G)$ are sorted in descending order by $\mathcal{R}(v) \times m_{v,u}$.

We compare RePBubLik+ also to two existing methods, ROV [26], and node2vec [31]. The *ROV* algorithm outputs a set of k edges to be added to G to minimize the controversy score (RWC) [29]. The RWC is a metric that characterizes how controversial a topic is by capturing how well separated the two colors are. ROV considers as candidates the edges between the high-degree vertices of each color [26, Algorithm 1]. These edges are sorted by descending impact on the graph controversy score, and the top-k edges are added to the graph. The objective of the comparison between ROV and RePBubLik+ is to verify whether an algorithm developed to minimize the RWC can be used to minimize the structural bias. *node2vec* is a graph embedding technique that encodes a network in a low-dimensional space retaining characteristics like the nodes' similarity [31]. The generation of the embedding is based on random walks. One of the main applications of node2vec is to employ the embedding as the feature space to train link recommendation algorithms. The goal of comparing node2vec to RePBubLik+ is to understand how the predictions of widely-used link recommendation algorithms affect the network's structural bias. In the experiments, we create for each network a 128-dimensional space, then we train a logistic regression (*avg.* AUC 85%) over these features, and we predict the existence probabilities of edges from $\mathcal{P}(G)$. We add to the graph the top k edges according to these probabilities.

Datasets. We create graphs obtained from *Wikipedia*, *Amazon*[2] and *PolBlogs*[3]. Table 1 shows the relevant statistics.

From *Wikipedia* we consider four bi-partitioned subgraphs related to controversial topics: *politics, abortion, guns* and *sociology* [45]. Each node in the graph is a page, and is assigned to one color according to Wikipedia's categorization. Directed edges denote links, and are weighted using Wikipedia's clickstream data.[4]

The *Amazon* dataset contains metadata about *books* [37]. Given two book categories, the vertices are all the items in those categories, colored accordingly. There is a directed edge (u, v) if v appears in the list of items similar to u. The edge is weighted by v's sales rank.[5] We built three graphs by considering pairs of the following categories: *Mathematics & Technology (MaTe), History of Technology & Military Science (MiHi),* and *Mathematics & Astronomy (MaAs).*

The *Political Blogs* dataset is a directed network of hyperlinks between weblogs on US politics [1]. Each node represents a blog and is colored according to its political leaning. Links between blogs

[2] https://snap.stanford.edu/data/amazon-meta.html
[3] http://www-personal.umich.edu/~mejn/netdata/
[4] https://dumps.wikimedia.org/other/clickstream/
[5] Amazon sales rank is a metric of the relationship among products within one category based on their sales performance. It expresses how well a product is selling relative to other products in the same category.

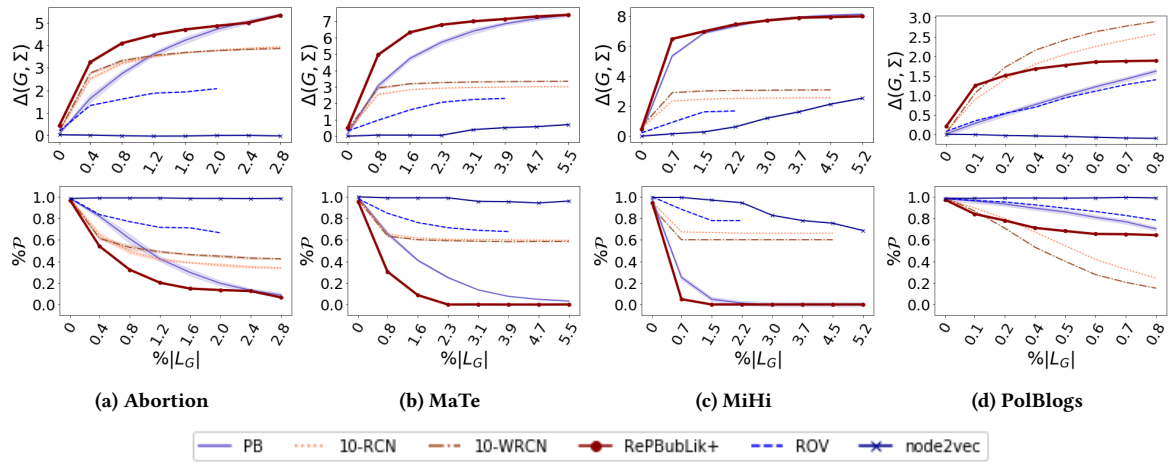


Figure 1: The first row shows the $\Delta(G, \Sigma)$ (y-axis) for increasing value of k, reported in terms of $\%\mathcal{L}_G$, the union of possible edges across $\mathcal{P}_C(G)$ and $\bar{C}$ for $C \in R, B$, (x-axis) for each algorithm. Higher values of Δ show more significant reduction of the structural bias. In the second row, we show the percentage of nodes that are still parochial, $\%\mathcal{P} = \frac{|\mathcal{P}(G)|-|\mathcal{P}(G_{\text{new}})|}{|\mathcal{P}(G)|}$ after k additions.

were automatically extracted from a crawl of the front page of the blog and represent the edges of the graph. Each edge (v, u) has weight proportional to the out-degree of v.

Wikipedia							
Topic	$\|R\|$	$\|B\|$	$\|E\|_{R\to B}$	$\|E\|_{B\to R}$	$\|E\|$	$\%\mathcal{P}_R(G)$	$\%\mathcal{P}_B(G)$
Abort.	208	413	80	170	1911	85.56	89.20
Guns	142	118	72	79	723	82.95	71.69
Pol.	10347	10129	17452	16484	141486	25.97	42.36
Sociol.	602	2283	284	192	10514	91.32	96.36
Amazon							
Topic	$\|R\|$	$\|B\|$	$\|E\|_{R\to B}$	$\|E\|_{B\to R}$	$\|E\|$	$\%\mathcal{P}_R(G)$	$\%\mathcal{P}_B(G)$
MaTe	827	566	25	42	675	90.91	79.63
MiHi	446	405	66	63	482	58.33	63.46
MaAs	827	294	11	6	680	97.31	95.15
PolBlogs							
Topic	$\|R\|$	$\|B\|$	$\|E\|_{R\to B}$	$\|E\|_{B\to R}$	$\|E\|$	$\%\mathcal{P}_R(G)$	$\%\mathcal{P}_B(G)$
Politics	545	488	902	781	17348	87.71	90.37

Table 1: Networks' statistics. The notation is consistent with the rest of the paper.

Setup. Given a network, we run RePBubLik and the other algorithms on that network for increasing values of K, with $K = 1, 2, 4, 6, \ldots, 400$ or 2000 for larger graphs (*Sociology* and *Politics*). These values of K represent only a small percentage of the set of possible edges to insert and correspond to the total number of edges to add to the graph. Once we set the value of K, accordingly, we allocate k_B and k_R of the K edge insertions to each color proportionally to the sum of the BRs of the parochial vertices in each color. In particular, we define $Y_C = \sum_{v\in\mathcal{P}_C(G)} \mathrm{B}_G^t(v)$, for $C \in R, B$, then $k_B = \left\lceil k\frac{Y_B}{Y_B+Y_R} \right\rceil$ and $k_R = K - k_B$. This allocation strategy is a simple but reasonable heuristic that ensures that more edges are added from nodes whose color is more parochial.

We assign the weight $m_{v,u} = 1/(d(v)+1)$ to the added edge (v, u), where $d(v)$ is the out-degree of v before the insertion, and then we re-normalize the weights of the other edges by multiplying each of them by $1 - m_{v,u}$. Furthermore, we set $r = 5$ and $b = 2$. Moreover, for the algorithms picking the top-N central nodes $N = 10$. To account for variability of the algorithm, we run them 10 times. The variance of the results is low, overall.

Reproducibility. The code for our experiments is available from https://github.com/CriMenghini/RePBubLik.

Experiment results. In Fig. 1, the plots in the first row show how the structural bias is affected by the insertion of an incrementally larger set of edges, while the ones on the second row show the reduction in the number of parochial nodes. Each curve in the plot illustrates the gain by a different algorithm. We can draw the following observations. (1) RePBubLik+ performs better than the baselines and the competitors, especially after the insertion of a few edges, as they obtain much larger gain with fewer insertions, i.e., the average BR of parochial nodes decreases faster requiring less modifications. (2) N-RCN, N-WRC, and ROV after a certain point become flat. (3) Overall, RePBubLik+ is the best algorithm. (4) The values of RePBubLik+ and PR converge, at different speed, to the same value when we add more edges. (5) node2vec, in the best cases, shows little improvement of the structural bias that, in the remaining cases, stays flat or even increases. We now explain these behaviours using the plots on the second row of Fig. 1.

(1) RePBubLik+ chooses edges that directly affect the BR of central nodes and, with a chain effect, the BR of nodes connected to them. More central are the nodes we attach the edges to, higher the structural bias drop is. In fact, it follows, as shown for all the

networks, that the addition of even small set of edges is very effective. Additionally, we observe that the structural bias reduction corresponds to a significant drop of the number of parochial nodes.

(2) N-RCN, N-WRC, and ROV attach edges only to a subset of $\mathcal{P}(G)$ and as k increases, so does the probability of adding multiple edges to the same nodes. These facts imply respectively that, especially on disconnected graphs (see MiHi in Fig. 1c), the addition of edges may affect few nodes, and that even the insertion of more edges does not modify the set of nodes on which the new edges have effect. Thus, the curves of N-RCN, N-WRCN and ROV reach an early saturation that expresses the scarce impact of subsequent edge additions. This explanation is confirmed by the percentage of parochial nodes, which does not decrease after the saturation point. Furthermore, the ROV shows a stepping behaviour due to it selecting edges between high-degree central nodes that minimize the RWC without imposing diversity constraints on nodes. And resulting in many selected edges being attached to the same node. Last, we see that on *Polblogs* the best algorithms are N-RCN, N-WRCN. This surprising superiority of the random approaches can be explained by the fact that *Polblogs* is a connected graph, thus edges added to the top-central nodes potentially affect all the nodes in $\mathcal{P}(G)$. Thus, even when N-RCN and N-WRCN add multiple new edges to the same set of nodes, Δ continues to increase.

(3) RePBubLik+ shows a consistent behaviour, indeed it increases the gain faster than other methods, requiring fewer insertions. The penalty factor η allows the algorithm to diversify the set of nodes to which the new edges attach, raising the chances of lowering the BR of a larger number of parochial nodes, thus increasing the gain. This feature is important especially on disconnected graphs, where the vertices in tiny connected components always have lower centrality compared to those in huge ones. More importantly, we observe that the size of $\mathcal{P}(G)$ is often reduced to 0: RePBubLik+ is able to "heal" all the bad vertices, and if we measured the structural bias on the obtained graph it would be zero.

(4) The variants of RePBubLik: RePBubLik+ and PR, pick edges from the same candidate set, thus the more edges they can pick, the more likely they choose edges with similar effect, thus the average parochial nodes' BR converges. This is the main explanation why the random algorithm performs so well.

(5) Generally, link recommendation algorithms tend to suggest edges between similar nodes. Node2vec captures this similarity through the nodes' neighborhood. In this context, graphs partitions have high within- and low between-density. Nodes in the same partition then lie close in the embedding space. Edges suggested by node2vec with high probability connect nodes close to each other in the embedding, which often are in the same partition. Thus, node2vec has a hard time reducing the structural bias, and in some cases increases it.

Due to space constraint we omit the presentation of the plots for *guns*, *sociology*, *politics* and *MaAs*, which show similar behaviour. But they can be found in the extend online version [32].

7 CONCLUSION

We presented RePBubLik, an algorithm that reduces the structural bias of a graph by adding k edges. Thanks to the monotonicity and submodularity of the objective function, RePBubLik is able to return a constant-factor approximation using a greedy approach based on a task-specific variant of the random walk closeness centrality. The results of our experimental evaluation show that the edge insertions suggested by RePBubLik result in a much quicker decrease of the structural bias than existing methods and reasonable baselines.

The functionality of RePBubLik relies on the existence of an oracle receiving the network and a page in it as input and outputting the transition probabilities of potentially added links to the input page. We leave the question of designing an algorithm which learns such probabilities from data as future direction of this work.

ACKNOWLEDGMENTS

Shahrzad Haddadan was supported by NSF Award CCF-1740741. Part of Cristina Menghini's work was done while visiting Brown University and is supported by the ERC Advanced Grant 788893 AMDROMA. Matteo Riondato is supported in part by National Science Foundation award IIS-2006765. Eli Upfal was supported in part by NSF awards RI-1813444, and CCF-1740741. We thank an anonymous reviewer for correcting one of our lemmas.

REFERENCES

[1] Lada A. Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (Chicago, Illinois) *(LinkKDD '05)*. Association for Computing Machinery, New York, NY, USA, 36–43. https://doi.org/10.1145/1134271.1134277

[2] Leman Akoglu. 2014. Quantifying political polarity based on bipartite opinion networks. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[3] Aris Anagnostopoulos, Luca Becchetti, Adriano Fazzone, Cristina Menghini, and Chris Schwiegelshohn. 2020. Principal Fairness: Removing Bias via Projections. arXiv:1905.13651 [cs.DS]

[4] Eugenio Angriman, Alexander van der Grinten, Aleksandar Bojchevski, Daniel Zügner, Stephan Günnemann, and Henning Meyerhenke. 2020. Group Centrality Maximization for Large-scale Graphs. In *2020 Proceedings of the Twenty-Second Workshop on Algorithm Engineering and Experiments (ALENEX)*.

[5] Francesca Arrigo and Michele Benzi. 2016. Edge modification criteria for enhancing the communicability of digraphs. *SIAM J. Matrix Anal. Appl.* 37, 1 (2016), 443–468.

[6] Francesca Arrigo and Michele Benzi. 2016. Updating and downdating techniques for optimizing network communicability. *SIAM Journal on Scientific Computing* 38, 1 (2016), B25–B49.

[7] Cigdem Aslay, Antonis Matakos, Esther Galbrun, and Aristides Gionis. 2018. Maximizing the Diversity of Exposure in a Social Network. In *2018 IEEE International Conference on Data Mining (ICDM)*. 863–868.

[8] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[9] Ruben Becker, Federico Corò, Gianlorenzo D'Angelo, and Hugo Gilbert. 2020. Balancing Spreads of Influence in a Social Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (2020), 3–10.

[10] Anna Ben-Hamou, Roberto I. Oliveira, and Yuval Peres. 2018. Estimating Graph Parameters via Random Walks with Restarts. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (New Orleans, Louisiana) *(SODA '18)*. Society for Industrial and Applied Mathematics, USA, 1702–1714.

[11] Seyla Benhabib. 1996. Toward a deliberative model of democratic legitimacy. In *Democracy and difference: Contesting the boundaries of the political*. Princeton University Press, Princeton, N.J., 67–94.

[12] Suman K. Bera and C. Seshadhri. 2020. How to Count Triangles, without Seeing the Whole Graph. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3394486.3403073

[13] Elisabetta Bergamini, Pierluigi Crescenzi, Gianlorenzo D'Angelo, Henning Meyerhenke, Lorenzo Severini, and Yllka Velaj. 2018. Improving the betweenness centrality of a node by adding links. *Journal of Experimental Algorithmics (JEA)* 23 (2018), 1–32.

[14] Hau Chan, Leman Akoglu, and Hanghang Tong. 2014. Make it or break it: Manipulating robustness in large networks. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 325–333.

[15] Flavio Chierichetti and Shahrzad Haddadan. 2018. On the Complexity of Sampling Vertices Uniformly from a Graph. In *45th International Colloquium on Automata, Languages, and Programming* (Prague, Czech Republic) *(ICALP 2018)*.

[16] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM.

[17] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Fifth international AAAI conference on weblogs and social media*.

[18] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the Echo Chamber: The Italian Vaccination Debate on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[19] Gianlorenzo D'Angelo, Martin Olsen, and Lorenzo Severini. 2019. Coverage centrality maximization in undirected networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 501–508.

[20] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. 2014. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 403–412.

[21] Anirban Dasgupta, Ravi Kumar, and Tamas Sarlos. 2014. On Estimating the Average Degree. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 795–806. https://doi.org/10.1145/2566486.2568019

[22] Erik D Demaine and Morteza Zadimoghaddam. 2010. Minimizing the diameter of a network using shortcut edges. In *Scandinavian Workshop on Algorithm Theory*. Springer, 420–431.

[23] Ioana Dumitriu, Prasad Tetali, and Peter Winkler. 2003. On Playing Golf with Two Balls. *SIAM J. Discrete Math.* 16 (2003), 604–615.

[24] Ronald Fagin, Anna Karlin, Jon Kleinberg, Prabhakar Raghavan, Sridhar Rajagopalan, Ronitt Rubinfeld, and Andrew Tomkins. 2001. Random Walks with "Back Buttons". *The Annals of Applied Probability* 11 (06 2001).

[25] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.

[26] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing Controversy by Connecting Opposing Views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*.

[27] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.

[28] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems*. 4663–4671.

[29] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* (2018).

[30] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. 257–260.

[31] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[32] Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. 2021. RePBubLik: Reducing the Polarized Bubble Radius with Link Insertions. *CoRR* abs/2101.04751 (2021). arXiv:2101.04751 https://arxiv.org/abs/2101.04751

[33] Daniel J. Isenberg. 1986. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology* 50, 6 (1986), 1141.

[34] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2016. Challenges of serendipity in recommender systems. In *WEBIST 2016: Proceedings of the 12th International conference on web information systems and technologies*.

[35] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*. 933–943.

[36] Rob LeFebvre. 2017. Obama Foundation taps social media to fight online echo chambers. (2017).

[37] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. 2007. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)* 1, 1 (2007), 5–es.

[38] Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 184–196.

[39] Q. Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2745–2754.

[40] Ahmad Mahmoody, Charalampos E Tsourakakis, and Eli Upfal. 2016. Scalable betweenness centrality maximization via sampling. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[41] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery* 31 (2017), 1480–1505.

[42] Antonis Matakos, Sijing Tu, and Aristides Gionis. 2020. Tell me something my friends do not know: diversity maximization in social networks. *Knowledge and Information Systems* 9 (2020), 3697–3726.

[43] Sourav Medya, Arlei Silva, Ambuj Singh, Prithwish Basu, and Ananthram Swami. 2018. Group centrality maximization via network design. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 126–134.

[44] C. Menghini, A. Anagnostopoulos, and E. Upfal. 2019. Wikipedia Polarization and Its Effects on Navigation Paths. In *2019 IEEE International Conference on Big Data (Big Data)*. 6154–6156.

[45] Cristina Menghini, Aris Anagnostopoulos, and Eli Upfal. 2020. Wikipedia's Network Bias on Controversial Topics. https://arxiv.org/abs/2007.08197

[46] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M. Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25, 3 (2015), 033114.

[47] Elchanan Mossel and Omer Tamuz. 2017. Opinion exchange dynamics. *Probability Surveys* 14 (2017), 155–204.

[48] Sean A Munson, Stephanie Y. Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[49] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. 2018. Minimizing Polarization and Disagreement in Social Networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*.

[50] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 957–970.

[51] Manos Papagelis, Francesco Bonchi, and Aristides Gionis. 2011. Suggesting ghost edges for a smaller world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2305–2308.

[52] Nikos Parotsidis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2015. Selecting shortcuts for a smaller world. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 28–36.

[53] Nikos Parotsidis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2016. Centrality-aware link recommendations. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 503–512.

[54] Senni Perumal, Prithwish Basu, and Ziyu Guan. 2013. Minimizing eccentricity in composite networks via constrained edge additions. In *MILCOM 2013-2013 IEEE Military Communications Conference*. 1894–1899.

[55] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*.

[56] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing Radicalization Pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. 131–141.

[57] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Hegemony in Social Media and the effect of recommendations. In *Companion Proceedings of The 2019 World Wide Web Conference*.

[58] Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. 2020. Seeding Network Influence in Biased Networks and the Benefits of Diversity. In *Proceedings of The Web Conference 2020*. ACM.

[59] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks. In *Proceedings of the 2018 World Wide Web Conference*. ACM Press.

[60] Cass R. Sunstein. 2002. The Law of Group Polarization. *Journal of Political Philosophy* 10, 2 (2002), 175–195.

[61] Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 245–254.

[62] Tomasz Wąs, Marcin Waniek, Talal Rahwan, and Tomasz Michalak. 2020. The Manipulability of Centrality Measures-An Axiomatic Approach. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1467–1475.

[63] Scott White and Padhraic Smyth. 2003. Algorithms for Estimating Relative Importance in Networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*. 266–275.

[64] An Zeng, Linyuan Lü, and Tao Zhou. 2012. Manipulating directed networks for better synchronization. *New Journal of Physics* 14, 8 (2012), 083006.

scientific reports

Check for updates

OPEN

Growing urban bicycle networks

Michael Szell[1,2,3]✉, Sayat Mimar[4], Tyler Perlman[4], Gourab Ghoshal[4] & Roberta Sinatra[1,2,3,5]

Cycling is a promising solution to unsustainable urban transport systems. However, prevailing bicycle network development follows a slow and piecewise process, without taking into account the structural complexity of transportation networks. Here we explore systematically the topological limitations of urban bicycle network development. For 62 cities we study different variations of growing a synthetic bicycle network between an arbitrary set of points routed on the urban street network. We find initially decreasing returns on investment until a critical threshold, posing fundamental consequences to sustainable urban planning: cities must invest into bicycle networks with the right growth strategy, and persistently, to surpass a critical mass. We also find pronounced overlaps of synthetically grown networks in cities with well-developed existing bicycle networks, showing that our model reflects reality. Growing networks from scratch makes our approach a generally applicable starting point for sustainable urban bicycle network planning with minimal data requirements.

Cities worldwide are scrambling for sustainable solutions to their inefficient, car-centric transport systems[1,2]. One promising, time-tested candidate is cycling. It is an efficient mode of sustainable urban transport that can account for the majority of intra-urban trips which are primarily short or medium-distance[3]. Cost-benefit analysis that accounts for health, pollution, and climate, reveals that in the EU alone cycling brings a yearly benefit worth € 24 billion while automobility costs society € 500 billion[4]. These insights provide further impetus for coordinated efforts to extend cycling infrastructure as one solution to the urban transport crisis and to effectively fight climate change[5,6]. Apart from being effective, this solution is also considerably more economic and wide-ranging than merely focusing on motor vehicle electrification[7–9].

In practice, however, bicycle infrastructure development struggles with a political inertia due to the deep-rooted complexity of car-dependence[10,11]: For example, Copenhagen took 100 years of political struggles to develop a functioning grid of protected on-street bicycle networks[12] that continues to be split into 300 disconnected components today[13]. Accordingly, the most developed, influential bicycle network planning guidelines, such as the Dutch CROW manual[14], acknowledge that building up bicycle networks happens typically through decades-long, piecewise refinements. Unfortunately, there is overwhelming scientific consensus that the possible exit scenarios from the planetary climate crisis compatible with the 1.5° goal are closing rapidly[15,16]. Given that transport is the most problematic sector[17] and that the majority of humanity is living in cities, making urban transport sustainable is therefore one of the most urgent societal issues[1,7,10,18]. Electric cars are a potential solution to exhaust pollution but come with the same unavoidable downsides as traditional cars concerning urban livability[2], space allocation[6], road safety[19], particulate matter pollution that is mainly caused by non-exhaust emissions[20], public health and equity[5,21], among others. In particular, a sole focus on electric vehicles is counterproductive and "active travel should be a cornerstone of sustainability strategies, policies and planning"[9]. Because of the fact that boosting active travel in cities has some of the highest potential to mitigate climate change and to improve public health[5,7], in this paper we focus on bicycle network development. While there has been historical political inertia in growing bicycle networks, the ongoing COVID-19 pandemic has prompted several cities to engage in successful accelerated network development, proving that such efforts are indeed possible[22,23] (apart from already existing examples of fast growth[24,25]). A systematic exploration of city-wide, comprehensive development strategies is therefore urgently needed.

Although the prevailing, piecewise application of qualitative policy guidelines in existing bicycle network planning[14,26] might have a good track record in e.g. Dutch cities and Copenhagen[12], this process lacks rigorous scrutiny: are the resulting networks truly optimal? Can such policies be replicated in other cities? And are there fundamental topological limitations for developing a bicycle network? Indeed, an evidence-based, scientific theory of bicycle network development is missing.

There is a growing academic literature on analyzing bicycle networks of specific cities, for instance Montreal[27], Seattle[28], or recent data-driven approaches for Bogota[29], London[30], or Berlin[31]. While such studies are invaluable in terms of local enhancements and data consolidation for a particular place, here we instead focus on a global

[1]NEtwoRks, Data, and Society (NERDS), IT University of Copenhagen, 2300 Copenhagen, Denmark. [2]Complexity Science Hub Vienna, 1080 Vienna, Austria. [3]ISI Foundation, 10126 Turin, Italy. [4]Department of Physics and Astronomy, University of Rochester, Rochester, NY 14627, USA. [5]Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, 1353 Copenhagen, Denmark. ✉email: misz@itu.dk

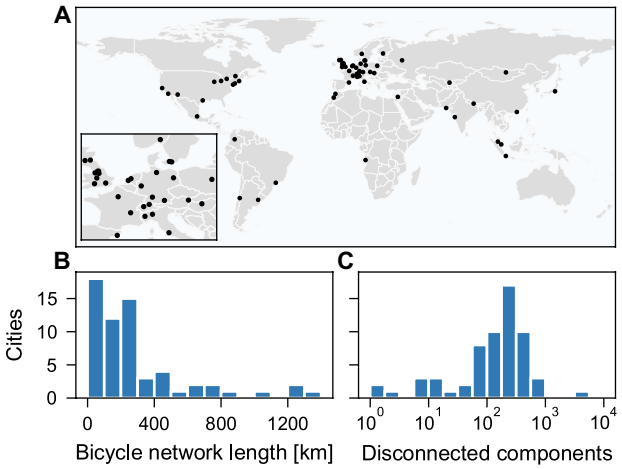


Figure 1. The state of existing bicycle networks. (**A**) We extract street networks from 62 cities covering different regions and cultures; many are considered modern and well developed. (**B**) The distribution of city-wide lengths of bicycle tracks indicates negligible existing cycling infrastructure that is also (**C**) split into hundreds of disconnected components. See more details in Supplementary Table 1. Map created with: https://github.com/mszell/bikenwgrowth (v.1.0.0).

analysis, in particular on the *fundamental topological limitations* of bicycle network development that are relevant for all urban environments, independent of the availability of traffic flow data[32]. This approach follows the idea of a *Science of Cities*[33] where we study the topological properties of bicycle networks that are *independent of place* using computational, quantitative methods of *Urban Data Science*[34].

The vast majority of cities on the planet has negligible infrastructure for safe cycling[6]. Indeed, urban transport infrastructure development worldwide has been heavily skewed towards automobiles since the twentieth century, today featuring well connected networks of streets for motorized vehicles[13]. Rather than uprooting the existing infrastructure and replacing it with an entirely new one—an economically infeasible strategy—we investigate how to retrofit existing streets into bicycle networks. Sacrificing specificity for generalizability, our formulation contains as a starting point two ingredients: the existing street network of a city, and an arbitrary set of seed points. With these minimal ingredients we explore different growth strategies that sequentially convert streets that were designed for only cars to streets that are safe for cycling[14,35]. Using the CROW manual as a key reference and inspiration[14], the objective of all explored strategies is to create *cohesive* networks, i.e. well connected networks that cover a large fraction of the city area (see "Materials and methods").

Across the realistic strategies we report a growth phase that initially leads to a diminishing trend of quality indicators, until a critical fraction of streets are converted, akin to a percolation transition observed in critical phenomena and also present in the growth of other forms of transportation infrastructure as well as patterns of traffic[29,32,36–39]. In other words, initial investments into building cycling-friendly infrastructure leads to diminishing returns on quality and efficiency until the emergence of a well-connected giant component. Once this threshold is reached, the quality improves dramatically, to an extent which depends on the specific growth strategy. We provide empirical evidence that the majority of cities effectively lie below the threshold which might be hindering further growth, implying fundamental consequences to sustainable urban planning policy: To be successful in developing well-connected bicycle networks, cities must invest with the right growth strategy, and *persistently*, to surpass a critical mass.

Results

The starting point for our analysis is the manually sampled street networks of 62 cities aiming to capture a diversity of cultural regions and a large range of populations, population densities, areas, and network lengths, selected from cities where there is relatively complete data available[41], see Fig. 1A and Supplementary Table 1. Here, links represent streets and nodes are street intersections. Being embedded in a metric space, these constitute planar graphs[42]. We downloaded and processed these networks from OpenStreetMap using OSMnx[43] (see "Materials and methods").

Although many of the covered cities are from well developed regions, we observe that they have negligible bicycle infrastructure, Fig. 1B. Additionally these are split into hundreds of disconnected components, Fig. 1C, which has previously prompted analysis of strategies to merge them[13]. Although such strategies make sense in cities with already well established bicycle infrastructure, they are less useful in most other cities. Further, they produce minimum spanning tree-like solutions that are economically attractive but lack resilience and cohesion (Fig. 2), and they potentially reinforce socioeconomic inequalities by connecting only already developed areas while ignoring under-developed ones[32], prompting us here to grow new networks from scratch instead. By resilience we mean a general level of fault-tolerance[44]: A resilient bicycle network should provide an acceptable level of service in the face of faults and challenges to normal operation, for example interruptions due to road works. The removal of a small fraction of links should not have a substantial impact on network metrics.

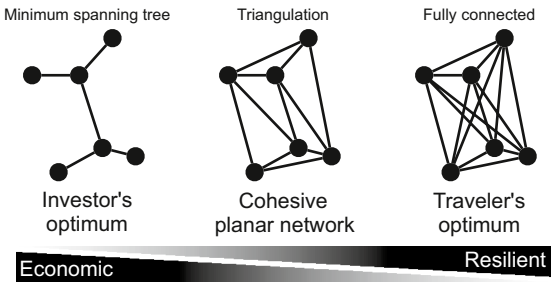


Figure 2. Optimal connected network solutions. Adapted from Ref.[40]. (Left) The investor's optimal strategy for a connected network is to invest as little as possible, minimizing total link length[13]. Its solution is a minimum spanning tree, maximally economic but minimally resilient with low directness, inadequate for travelers. (Right) The traveler's optimum connects all node pairs creating all direct routes. This solution is minimally economic, maximally resilient and direct, inadequate for investors. It also has crossing links and is therefore not a planar network. (Center) A both economic and resilient, as well as cohesive planar network solution in-between is the triangulation. In particular the minimum weight triangulation, approximated by the greedy triangulation, minimizes investment.

Growing bicycle networks from scratch. Our process of growing synthetic bicycle networks consists of four steps, Fig. 3, starting with the street network and the seed points. For an intuitive, interactive exploration see https://growbike.net.

Step (1) Seed points An arbitrary set of seed points is snapped to the intersections of the street network. We investigate two versions of seed points: (i) arranged on a grid, and (ii) rail stations. Generally these seeds could have arbitrary coordinates, but in the CROW manual's context of origin-destination links[14] they could represent points of interests such as district centers, shopping areas, schools, etc.
Step (2) Greedy triangulation All pairs of seed nodes are ordered by route distance and connected stepwise as the crow flies. A link is added only if it does not cross an existing link. This greedy triangulation is an easily computable proxy for the NP-hard minimum weight triangulation[45]. It creates an approximatively shortest and locally dense planar network[46], and a connected, cohesive, and resilient network solution minimizing investment, therefore satisfying both traveler and investor demands, Fig. 2.
Step (3) Order by growth strategy Each of three growth strategies is used to order the greedy triangulation links from the strategy's 0-quantile (empty graph) to its 1-quantile (full triangulation), resulting in a sequence of growth stages. To study this growth process in a high enough resolution we split the growth quantiles into 40 parts $q = 0.025, 0.05, \ldots, 0.975, 1$. The three strategies are:

1. *Betweenness*—orders by the number of shortest paths that go through a link. It can be interpreted as the simplest proxy for traffic flow (assuming uniform traffic demand between all pairs of nodes). Thus, growing by betweenness is an approach that aims to prioritize flow.
2. *Closeness*—starts with the "most central" node, i.e. the node that is closest to all other nodes. From this seed, the network is built up by connecting the most central adjacent nodes. This approach is the most local approach possible and leads to a linear expansion of a dense as possible network from the topological city center.
3. *Random*—adds links randomly and is used as a baseline. This strategy is not just a theoretical null model but well resembling how cities build their bicycle networks in practice, as we discuss later.

Step (4) Route on street network The abstract links in the 40 stages are made concrete: They are routed on the street network. These synthetic bicycle networks are then analyzed for all 62 cities.

Different growth strategies optimize different quality metrics. We measure several network metrics to assess the quality of the synthetically grown networks and to compare them with existing bicycle networks. These metrics are: length L, length L_{LCC} of the largest connected component (LCC), coverage C, seed point coverage C_{seed}, directness D, number of connected components Γ, global efficiency E_{glob}, local efficiency E_{loc}. We define coverage as the area of all grown structures endowed with a 500 m-buffer, see light blue areas in Fig 6B for an illustration. Directness measures the average ratios of euclidian distances versus shortest path distances on the network, while global efficiency provides a similar measure that accounts for disconnected components[47]. See "Materials and methods" for technical details.

We first investigate how these quality metrics change throughout the growth process averaged over all cities, Fig. 4. The three thick curves show the change of the metrics with growth following the three strategies (betweenness, closeness, or random) for grid seeds. Similar results hold for rail stations, see Supplementary Figs. 5–7. By construction all curves reach exactly the same point at the 1-quantile (full triangulation), but their development differs substantially before that. The minimum spanning tree (MST) solution is depicted as a baseline (grey dotted

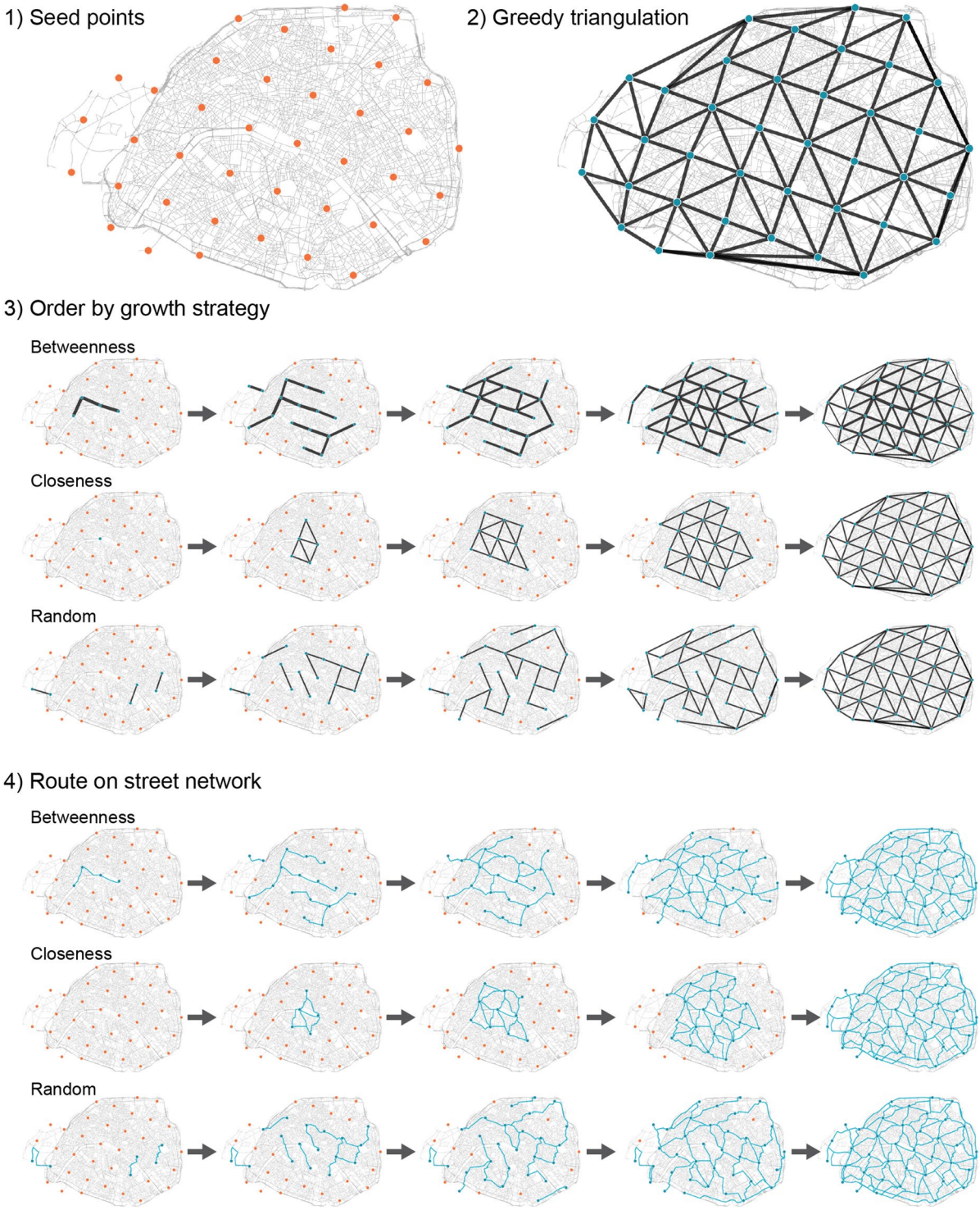


Figure 3. Growing bicycle networks. Explorable interactively at: https://growbike.net. Illustrated here for Paris. Step (1) Seed points: A set of seed points (orange dots) is snapped to the intersections of the street network. Shown are grid points, alternatively we investigated rail stations. Step (2) Greedy triangulation: The seeds are ordered by route distance and connected stepwise without link crossings. Reached seeds are colored blue. Step (3) Order by growth strategy: One of three growth strategies (betweenness, closeness, random) is used to order the triangulation links from the strategy's 0-quantile (empty graph) to its 1-quantile (full triangulation), resulting in 40 growth stages. Shown are the five quantiles $q = 0.025, 0.125, 0.25, 0.5, 1$. Step 4) Route on street network: The links in the growth stages are routed on the street network. These synthetic bicycle networks are then analyzed for all 62 cities. Maps created with: https://github.com/mszell/bikenwgrowth (v.1.0.0).

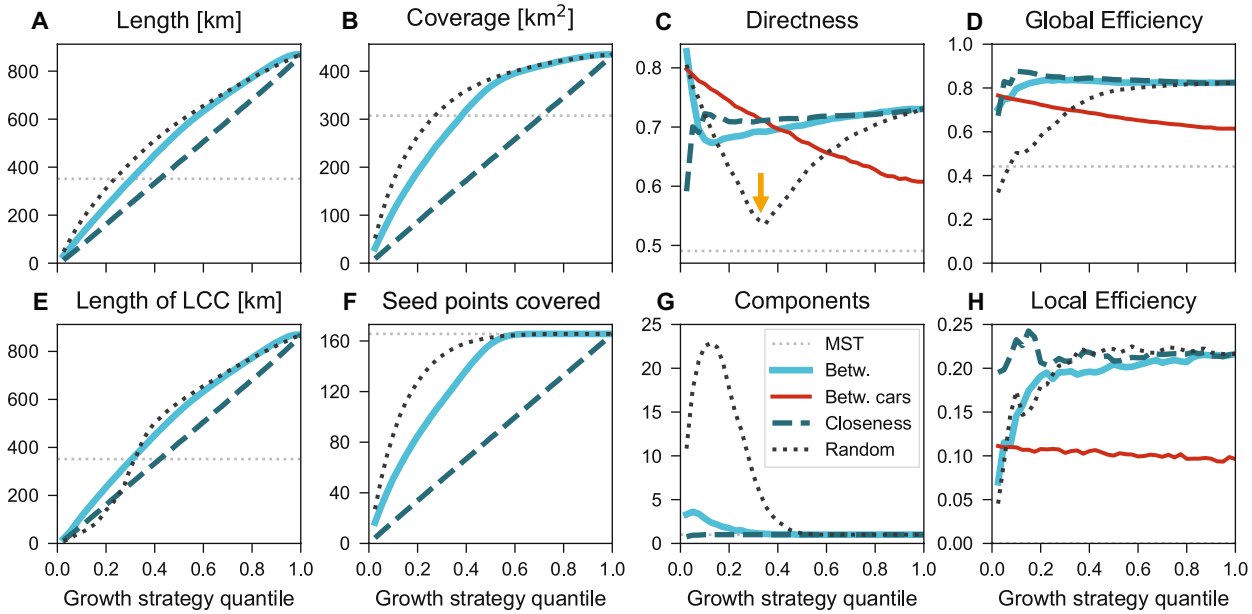


Figure 4. Different growth strategies optimize different network quality metrics. The three thick curves show the changes of network metrics with growth following three strategies (betweenness, closeness, or random) averaged over all 62 cities for grid seeds. By construction all curves arrive at the same endpoint, but they develop distinctly before that. For rail seeds and individual cities see Supplementary Figs. 5–10. Red curves show the car network's simultaneous decrease of quality metrics if a five times decrease of speed limits is assumed for cars along all affected streets. Grey dotted lines show metrics for the minimum spanning tree (MST) that connects all seeds with minimal investment. Growth of (**A**) length, (**B**) coverage, (**C**) directness, (**D**) global efficiency, (**E**) length of LCC, (**F**) seed points covered, (**G**) connected components, (**H**) local efficiency. The yellow arrow highlights the substantial dip in directness until the critical threshold which is more pronounced for random growth than for betweenness growth.

lines). This is the most economic connected solution that reaches all seeds, see Fig. 2; therefore any connected solution that reaches all seeds must be at least as long as the MST.

From Fig. 4A we observe that length grows linearly for closeness and slightly faster for the other strategies because closeness prioritizes close links which typically have similar length, while betweenness and random growth selects distant links earlier. Random growth adds single links scattered randomly across the city and therefore has the fastest growth of coverage, Fig. 4B, followed by the betweenness-based strategy, while closeness leads to a linear growth. Directness, Fig. 4C, displays a large dip for random growth, from $D \approx 0.8$ down to $D \approx 0.53$ at the 0.345-quantile, and a smaller dip from $D \approx 0.83$ to $D \approx 0.68$ for betweenness growth at the 0.1-quantile. Directness starts lower for closeness growth, around $D \approx 0.59$ but quickly overtakes the other strategies at the 0.05-quantile. Global efficiency, Fig. 4D, starts at a high level, around $E_{\text{glob}} \approx 0.7$, and grows slightly until $E_{\text{glob}} \approx 0.82$ for both betweenness and closeness. Random growth starts instead much lower, around $E_{\text{glob}} \approx 0.33$. The length of the LCC, Fig. 4E, is almost identical as the growth of length for betweenness and closeness because the LCC makes up most of the network here. However, the LCC in random growth has a sigmoid growth pattern as it takes longer for the components to connect, Fig. 4G. Coverage of seeds, Fig. 4F, is similar to coverage but more pronounced for random and betweenness growth. On average all seeds are covered before the 0.6-quantile. Finally, local efficiency, Fig. 4H, is steady around $E_{\text{loc}} \approx 0.22$ for closeness, but grows fast for both betweenness and random growth from around $E_{\text{loc}} \approx 0.05$.

To summarize, the different growth strategies optimize different quality metrics and come with different tradeoffs: (1) Use betweenness growth for fast coverage, intermediate connectedness and directness, and low local efficiency. (2) Use closeness growth for optimal connectedness and local efficiency but slow coverage. (3) Use random growth for fastest coverage but low directness, connectedness, and efficiency.

Network consolidation and non-monotonic gains in quality. The dips observed in directness, see yellow arrow in Fig. 4C for random growth, are akin to a phase transition in a percolation process from a disconnected set of components to a sudden emergence of a giant connected component, as known for e.g. random Erdős–Rényi networks[36]. Similar transitions have been observed in generalized network growth[48,49], including in various random spatial networks[42] and sidewalk networks[39], and similar flavors of bicycle network growth[29]. Figure 5A and B illustrates this consolidation process for individual cities: Links are added one by one, growing the largest connected component until a critical threshold at the curve's minimum (at $q_B = 0.1$ for Boston), at which the largest connected component consolidates the majority of the network and starts forming cycles that in turn increase directness. Because connectedness increases around the critical threshold, the evolution of connected components is inverse to the evolution of directness, Fig. 4G. While the global efficiency averaged over all

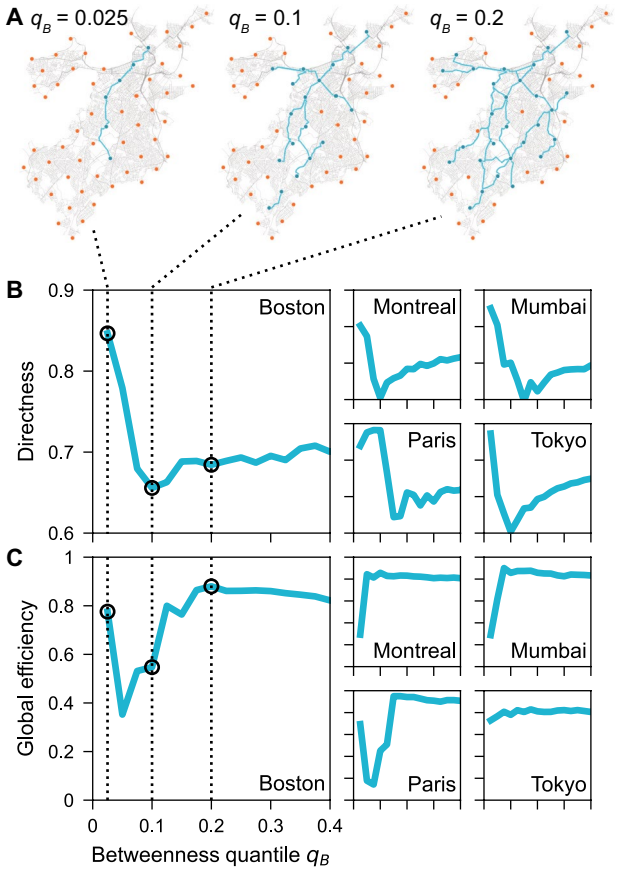


Figure 5. Network consolidation: Bicycle network growth has a dip of decreasing directness. (**A**) Three early stages of betweenness growth in Boston. (**B**) Directness sharply decreases initially due to tree-like growth (compare $q_B = 0.025$ and $q_B = 0.1$ for Boston). Once directness has reached a minimum ($q_B = 0.1$), it starts growing slowly due to the appearance of cycles ($q_B = 0.2$). The process is similar for the other cities (shown here for Montreal, Mumbai, Paris, Tokyo) and also holds for random growth, see Supplementary Figs. 5, 7, 8, 10. (**C**) We find mixed results for global efficiency: Mumbai and Montreal display a single jump, Tokyo is flat, while Boston and Paris shown an initial dip before increasing. Maps created with: https://github.com/mszell/bikenwgrowth (v.1.0.0).

cities shows an initial increase followed by saturation, see Fig. 4D, we find mixed trends at the level of individual cities: Mumbai and Montreal track the average trend, Tokyo has a flat global efficiency, while Boston and Paris show a dip before the critical threshold is reached with rapid gains thereafter (Fig. 5C).

This network consolidation has important implications for policy and planning. The point at which the transition happens represents substantial investments into building the network. Stopping investments and growth *before* this point leads to a net loss in investment as measured by infrastructure quality. Indeed, pushing past this threshold leads to substantive gains.

The effect of bicycle network growth on the street network. While it is beneficial for a city to grow its bicycle network, it is important to ask how this growth affects the network of streets used by cars. The magnitude of this effect depends not only on the network topology, but also on the concrete bicycle infrastructure being implemented: shared spaces, unprotected cycle lanes, protected cycle tracks, bicycle streets, their width, and so on. To consider these factors, leading bicycle planning manuals consider a plethora of local variables[14,50], for example road category, speed limit, volume of the motorized traffic, or car parking facilities. Therefore, to be conservative in our estimations, here we consider the strongest possible effect of new bicycle infrastructure on streets apart from complete replacement: We assume that all infrastructure would be built, for example, as a child-friendly "fietsstraat" or living street, i.e. as a shared traffic space where cyclists and pedestrians have priority and cars are tolerated to pass through in walking speed[14]. This assumption roughly translates to a reduction of speed limits for cars along the affected road sections by a factor of 5, for example from 50 km/h to 10 km/h or from 30 km/h to 6 km/h. In our technical calculations we implemented a computational equivalent to this speed reduction—an increase of the affected road section lengths by a factor of 5[14]. So, for example, for calculating directness along an affected road section, a street segment of length 100 m would then count as being 500 m long.

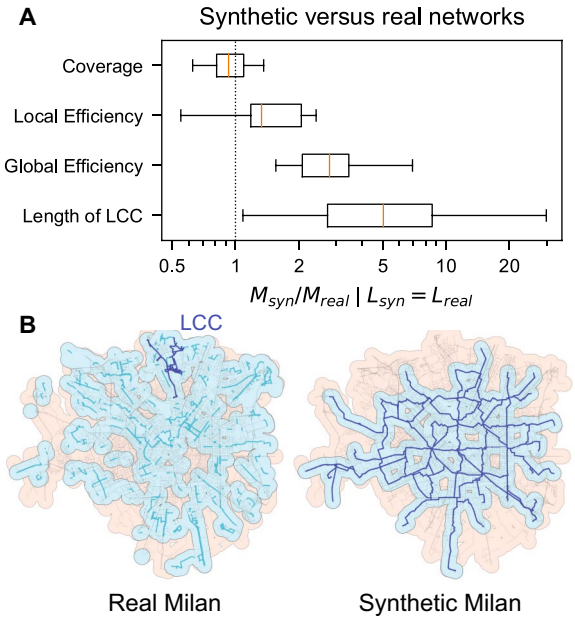


Figure 6. Synthetic bicycle networks perform several times better than existing ones. (**A**) We plot the distributions (over cities) of the ratios $M_{\text{syn}}/M_{\text{real}}$ between network metrics of synthetic and existing topologies fixed at same length ($L_{\text{syn}} = L_{\text{real}}$), for betweenness growth and grid seeds (for all other growths see Supplementary Fig. 2). Synthetic networks have on average 5 times larger LCCs, 3 times the global efficiency, and higher local efficiency. Existing networks only tend to have better coverage because they are more scattered. (**B**) Illustration of high coverage (light blue area) due to extreme scattering and low length of LCC (dark blue sub-network) for Milan's existing bicycle network, versus its synthetic version at same length (185 km at $q_B = 0.425$). The LCC for synthetic Milan is the whole network. Maps created with: https://github.com/mszell/bikenwgrowth (v.1.0.0).

Given this strong constraint, we find the metric that is most affected is directness. It decreases approximately linearly with the bicycle network's growth, from $D \approx 0.8$ to $D \approx 0.6$, as the network is grown using the betweenness strategy (red curve in Fig. 4C). In other words, in the absence of any bicycle infrastructure, car-routes deviate by around 25% from the Euclidean distance between any origin-destination, whereas once the full bicycle infrastructure is established, this increases to around 66%. At around the 0.4 betweenness quantile, the directness of the bicycle network exceeds that of the car network. The global efficiency decreases from around $E_{\text{glob}} \approx 0.75$ to $E_{\text{glob}} \approx 0.6$, (red curve in Fig. 4D), while the local efficiency decreases negligibly from $E_{\text{loc}} \approx 0.11$ to $E_{\text{loc}} \approx 0.10$ (red curve in Fig. 4H). Growing the bicycle network has no effect on the length and coverage of the automobile network, given that cars can still access all points on the street network, albeit in a longer time than they would without the bicycle infrastructure. We find almost identical behavior for the closeness and random growth strategies (Supplementary Fig.4).

One of the effects of modifying the street infrastructure is the redistribution of load on the street intersections, measured by the betweenness centrality. It has been shown that while the global distribution of the betweenness centrality remains unchanged due to change in density of streets, the spatial distribution and clustering of the high betweenness nodes tend to change, thus redistributing areas of higher traffic[51]. Two measures to quantify this effect are the spatial clustering and the anisotropy of the high betweenness nodes (see "Materials and methods"). We find a slight increase (around 5%) in spatial clustering and anisotropy for nodes in the 90th percentile of betweenness values but the effect is marginal (Supplementary Fig. 4).

Comparing synthetic with existing network metrics. Although the growth processes described here are somewhat artificial, given the lack of accounting for practical limitations of bicycle network design—street width, incline, or political feasibility for instance—it is nevertheless prudent to compare the synthetic network with existing bicycle networks to gauge their general correspondence. To have a fair comparison in terms of length (which is a proxy for cost), we first select all cities that have a protected bicycle network with shorter length L_{real} than the fully grown synthetic network L_{syn} (42 out of 62 cities), and for each of them we fix the growth quantile where the synthetic length is equal to the real length, $L_{\text{syn}} = L_{\text{real}}$. Given this set of bicycle network pairs—real versus synthetic at same length—we then measure the ratio $M_{\text{syn}}/M_{\text{real}}$ between the synthetic quality metric M_{syn} and the quality metric of the existing infrastructure M_{real}. The results for the metrics of coverage, local efficiency, global efficiency, and length of LCC are reported in Fig. 6A.

We find that synthetic networks have on average 5 times larger LCCs, 3 times the global efficiency, and higher local efficiency. Existing networks only tend to have better coverage because they are more scattered, as illustrated in Fig. 6B for Milan which has 230 disconnected components. Milan's scattered network provides an

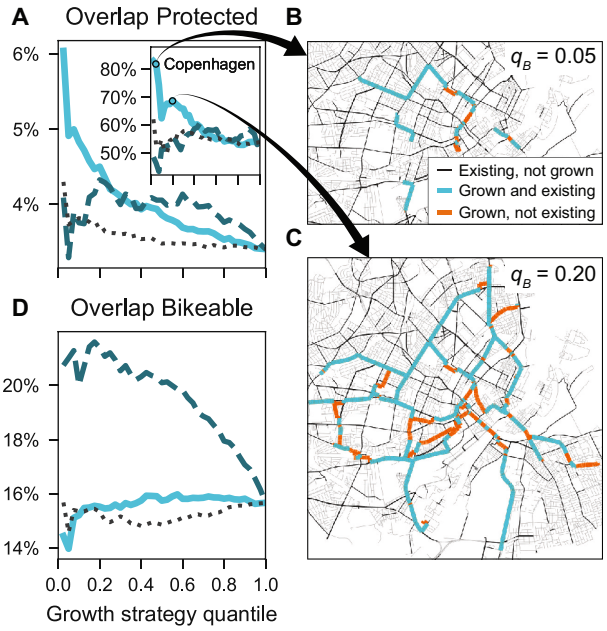


Figure 7. First stages of synthetic growth recreate existing networks. Shown are results for rail station seeds averaged over all cities. Same legend as Fig. 4. (**A**) Growth by betweenness starts with high, then decreasing overlap with existing protected bicycle infrastructure. Inset: The effect is especially strong in cities with well developed on-street bicycle networks such as Copenhagen. Here the growth algorithm starts with over 80% overlap. (**B**) Map of this high overlap in Copenhagen at the quantiles $q_B = 0.05$ and (**C**) $q_B = 0.20$. (**D**) The overlap with bikeable infrastructure has a notable effect only for growth by closeness due to traffic-calmed city centers: With increasing distance from the city center, overlap falls. Maps created with: https://github.com/mszell/bikenwgrowth (v.1.0.0).

important lesson: Mere measures of total length or coverage are misleading when it comes to an efficient and safe infrastructure if the network is not well connected. Instead, if city planners were to develop and implement bicycle networks holistically, considering a city-wide rather than piece-wise local approach, much higher quality infrastructure could be derived to the benefit of the residents. Many examples such as Dutch cities, Seville, or Paris have already proven that this is indeed a realistic approach[24,25,52].

For completeness we also compare our results to the closeness and random growth approaches, Supplementary Fig. 2. For closeness we find almost the same result as for betweenness, only with notably worse coverage which is to be expected given how closeness grows the covered area as slowly as possible. For the random growth approach we find the same coverage as existing infrastructure, and around 2 times the global efficiency and length of LCC. At first blush, this implies that even a naive random growth strategy can perform better than existing ones. However, this could be due to a number of reasons: For instance, in the random growth process described here, segments are added over at least 1.7 km in each step, whereas in real cities, segments are added in a more scattered fashion and in varying lengths. Further, cities can have non-negligible off-street bicycle tracks, for example through parks, a feature not considered in our analysis.

Comparing synthetic with existing network overlaps. To gain a better understanding into how the synthetically grown parts compare to existing infrastructure, and thus the extent to which the growth models approximate reality, we measure the percent overlap of synthetic infrastructure with existing bicycle infrastructure. Figure 7A and B report for rail station seeds the average overlaps for protected cycle tracks and for bikeable infrastructure respectively, where bikeable infrastructure is defined as the union of protected tracks and streets with speed limits $\leq$ 30 km/h (see "Materials and methods"). Results are qualitatively similar for grid seeds, see Supplementary Figs. 8–10.

For protected infrastructure, Fig. 7A, we find that growth by betweenness (solid line) starts with around 6% overlap on average, then decreases fast reaching 3.5%. We find a similar behavior for random growth (dotted line) but with a smaller effect. We find no clear effect for growth by closeness (dashed line). These observations suggest that cities take into account flow (betweenness) when building their cycling infrastructure, and that rail stations play some role—otherwise there would be no effect in the random growth. The betweenness overlap effect is especially strong in Copenhagen, Fig. 7A inset, which has a well-developed, cohesive on-street bicycle network. Figure 7B shows the synthetic growth stage at the second step, $q_B = 0.05$. Remarkably, at this step over 80% of the links suggested by the synthetic network model do already exist in reality. Even at $q_B = 0.20$ there is almost 70% overlap, Fig. 7C. These values are far higher than expected by chance: Given the length of Copenhagen's on-street cycle tracks we would expect only at most 24% overlap in a random link placement.

The overlap for bikeable infrastructure looks different, Fig. 7D. Here, there is no clear effect for betweenness and random growth, but a very clear effect for closeness, which starts with high values (above 20% on average), falling slowly to 16%. This observation is consistent with cities preferentially installing low speed limit areas in their city centers.

Discussion

We grew synthetic bicycle networks in 62 cities following three different growth strategies all aiming to generate a cohesive network, i.e. a network that is well connected and covers a large fraction of the city area. Studying the resulting networks we found a consistent critical threshold affecting directness in all cities and global efficiency in some of them, for the two most realistic strategies of growth by betweenness and random growth. This sudden network consolidation therefore has a fundamental policy implication: To grow bicycle networks successfully, cities must invest into them *persistently*, to surpass short-term deficiencies until a critical mass of bicycle infrastructure has been built up. Further, from a topological perspective, cities should avoid traditional "random growth-like" strategies that follow local, stepwise refinement. Such strategies substantially shift the critical threshold, thereby hold up the development of a functional cycling infrastructure which could fuel adversarial objections to bicycle network expansions along the lines of "We already built many bike tracks but nobody is using them, so why build more?" As we have shown, *it is not a network's length that matters but how you grow it*.

Our main result focuses on directness because it is the most important metric for bicycle network planning apart from connectivity: It is the key metric to quantify network quality[14], and it is the best predictor or quantifiable policy aspect for adoption of cycling[53,54]. To ensure that our results are robust to other possible definitions of directness, we compared our main result, see Fig. 4C, using four different definitions, see "Materials and methods" and Supplementary Fig. 3. Numerical values vary only insignificantly, all results are qualitatively identical for each definition, thereby establishing robustness.

By comparing metrics and overlaps of synthetic with existing networks, we gained an insight into the realism of our models. Our observations suggest that growth processes of existing protected bicycle networks contain a strong random ingredient and a detectable consideration for flow (betweenness) and rail stations. The random ingredient can be explained by the traditionally slow-paced urban planning processes arising from political inertia[11,26]. Unfortunately, the random strategy is also the slowest in terms of network consolidation: It needs at least three times the investments than the betweenness strategy to reach the critical threshold. The rail effect can be explained by transit-oriented development efforts, where bicycle facilities are planned close to transit lines[30,55]. The remarkably high overlap with Copenhagen's well developed network suggests that our models could also be adapted to identify "missing links" in existing bicycle networks[56,57].

Although the emergence of a giant component in network growth could have been anticipated with network science expertise, our results are not trivial: (1) This crucial insight is missing in the bicycle network planning manuals that practitioners use[14], (2) the different pros and cons of growth strategies have not been studied nor quantified before, (3) the policy dimension shows that reports on lengths and functionalities of under-developed bicycle networks must be scrutinized in an evidence-based way and take into account network structure. Only by being global and minimalistic, deliberately ignoring second order effects, does our approach uncover fundamental topological limitations of bicycle network growth independent of place. At the same time our results must be treated as *statistical solutions*. By no means do they suggest concrete recommendations for new bicycle facilities, as a vast array of local idiosyncrasies (second order effects) would need to be accounted for[14,50], including: road category, speed limit, volume of motorized traffic, or aspects of comfort[58]. Despite the importance of these aspects, a transport network's geometry is its most fundamental limitation[59] which is the reason we explored it first. Although our approach here is not yet aiming to provide concrete urban design solutions, it could be useful for planning purposes for easily generating an initial vision of a cohesive bicycle network—to be refined subsequently[56]. By publishing all our code as open source we facilitate such future refinements. Our minimal requirements on data are a deliberate limitation we impose for our framework to be applicable to data-scarce environments and thus to a large part of the planet[41]: no lane widths, inclines, traffic flows, etc. are needed to optimize network topology.

The studied alternative approach of starting from rail station seeds instead from grid seeds seems reasonable, however care has to be taken to not amplify existing biases that are well-documented in the transport planning profession[21,60,61]: For example, planning bicycle infrastructure only along metro stations that were built following elitist or racist biases would reinforce them, neglecting under-served regions and their inhabitants even further. The strength of our seed point approach lies in its arbitrariness that can bypass such issues: Grid seeds implement equal coverage and could be a starting point, to be refined carefully with e.g. population density, traffic demand models, or flow data[32,56,62]. The biggest limitation of our approach is the sole focus on retrofitting street networks for safe cycling. This approach has some issues because it only considers on-street but no off-street bicycle infrastructure. We discuss the technical details of this limitation, mostly relevant for concrete bicycle network planning in low urban density, in Supplementary Note 1, concluding that future research on bicycle network growth should consider off-street solutions wherever possible.

Finally we discuss the effect of growing bicycle networks on limiting street networks for car traffic. Our flow analysis detects no substantial change of choke points. To be fair, this analysis is static and does not account for possibly nonlinear dynamic congestion effects which could be studied in arbitrary detail and precision. However, the state of the art in sustainable travel planning and systems design is clear that such short-term dynamics predictions are overtrumped by long-term behavioral effects[63–67]: Induced demand posits that the development of a functional cycling infrastructure will generally drive a modal shift towards cycling— for latest evidence see, e.g. Refs.[23,24]—while the reclamation of ineffectively used automobile space will naturally lead to disappearing traffic. Therefore, the OECD recommends to replace the outdated "predict and provide" planning paradigm with

the vision-led "decide and provide" principle[64,65]. Our research follows this principle by prioritizing planning for access and the latent demand for cycling[68,69] through a cohesive network, rather than optimizing hard to forecast flow dynamics that are trumped by stronger equilibrium effects in the long term.

Concerning the change from directness $D \approx 0.8$ to $D \approx 0.6$, it is unclear whether to interpret it as substantial or inconsequential. Following considerations of long-term systems design as above, we deem it more important to discuss whether a small or a large change is *desired*. There are arguments for both sides: From the perspective of car-dependent transport planning the change should be small to not disrupt the existing system too abruptly[10,66,70]. From the perspective of sustainability, human-centric urban planning, and climate research, the change should be large to boost efficient bicycle transport, livable cities, and to fight climate change effectively. Indeed, the CROW manual states that directness should be higher for cyclists than for cars[14]. On top of that it could be argued that our eurocentric Copenhagen-style model of building a relatively sparse sub-network for cyclists is not going far enough, or that it could be out of place in other socio-cultural or land-use contexts[71–73]. For example, it could be inverted into a Barcelona-style model where dense patches of living streets – Superblocks – are built within a sparse sub-network of automobile arterials[66,74]. In any case, resistance to such ideas needs to be anticipated[5,70], requiring vigorous policy making and a well-informed civil society following leading examples such as the Netherlands[11,52]. Sustainability science provides overwhelming evidence for the societal benefits of following such persistent implementations, facilitating the transition to cities with sustainable transport systems to counteract climate change effectively, and providing extraordinary benefits to public health and urban livability[2,9,18,19,67,75].

Summarizing limitations and future work, we call for network development models that combine both the long-term goal of a cohesive, accessibility-focused network as we do here, and the use of empirical, place-specific or street-level data for refinements[56], while being critical of flow-optimizing engineering approaches[66]. On a policy level, more research is needed into understanding socio-technical processes to overcome political inertia[10,72,73]. Finally, let us answer the questions posed in the beginning. Are bicycle networks of existing cities optimal?—Our example of Milan has shown that in general, they are not, or that they are built in a too disconnected way. However, when it comes to well developed cities like Copenhagen, we find—despite many still outstanding gaps[57]—higher than expected overlap in the first growth stages, showing signs of an optimization process. Can optimal growth policies be replicated in other cities?—Yes, the technical solutions exist, and the scale of investment is mostly a matter of political will as we can see from the Netherlands, Sevilla, or Paris[24,25,52]. And are there fundamental topological limitations for developing a bicycle network? Yes, there is a critical threshold to overcome until a functional bicycle network emerges. Because of this threshold and its dependence on the growth strategy, our practical recommendations are to concentrate investments as early as possible, and to grow for the whole city instead of piece-wise.

Materials and methods

Network data and growth. *Infrastructure networks.* We downloaded existing street and bicycle networks for 62 cities from OpenStreetMap (OSM) on 2021-02-26 using OSMnx[43]. For each city, three networks were downloaded: Street network, protected bicycle network, bikeable network. Each node is an intersection, each link is a connection between two intersections. A protected bicycle network is the union of all OSM data structures that encode protected bicycle infrastructure, both on-street and off-street. Following the cycling safety literature, we consider only protected bicycle networks in our main analysis because safe cycling in general conditions is only ensured through physical separation from vehicular traffic[6,14,35,50,52] We also consider for additional analysis the "bikeable" network, which is the union of a protected bicycle network and all streets with speed limits ≤ 30 km/h or ≤ 20 m/h (including living streets). In special conditions such street segments can be considered safe for cycling, but not in general[14,35]; safety is a complex topic requiring a deep discussion of a multitude of variables[19], therefore we consider it outside the scope of this work. OSM data has been generally found to be of high quality and completeness[41,76], but multicity studies using bicycle infrastructure data such as ours could potentially suffer from some labeling inconsistencies, especially for less common types of bicycle infrastructure[77].

Seed points. Rail station seeds consist of all railway and metro stations. A few of the considered cities do not have rail stations. For creating grid seeds, we created grid points at a distance of 1707 m, ensuring a tolerable average distance of 167 m (2 mins walking) over the whole city to the triangulated network in the worst case, see Supplementary Note 2. We then rotated this grid to align it with the city's most common street bearing[78], and snapped the grid points to the closest street network intersections within a 500 m tolerance. The rotation is mostly important for US cities that have a grid-like street network, e.g. Manhattan, for creating straight routes.

Greedy triangulation. The greedy triangulation orders all pairs of nodes by route distance and connects them stepwise as the crow flies. A link is added only if it does not cross an existing link. This triangulation is a $O(N \log N)$ computable proxy for the NP-hard minimum weight triangulation[45] with an approximation ratio of $\Theta(\sqrt{N})$[79]. The greedy triangulation is fast and solvable for any set of nodes. Computing a quadrangular grid, as suggested by the CROW manual[14], or a quadrangulation, is in general not possible for arbitrary sets of nodes and also computationally less feasible[80].

Growth strategy: betweenness centrality. This is a path-based measure that computes the fraction of paths passing through a given node i[81],

$$C_B(i) = \frac{1}{N} \sum_{s \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{1}$$

where σ_{st} is the number of shortest paths going from nodes s to t and $\sigma_{st}(i)$ is the number of these paths that go through i.

Growth strategy: closeness centrality. Measures the total length of the shortest paths from a node i to all other nodes in the network[82],

$$C_C(i) = \frac{N-1}{\sum_{i \neq j} d(i,j)} \tag{2}$$

Network metrics. *Cohesion.* The CROW manual[14] describes qualitatively what it means for a network to be *cohesive*: a "combination of grid size and interconnection". It states that this is the most elementary requirement for a bicycle network but without a rigorous definition. We interpret this concept as having both high connectedness (few disconnected components) and coverage, see below. A cohesive network should also be resilient, see below, which excludes pathological cases like the minimum spanning tree.

Coverage. We measure spatial coverage of the network as the union of the ε-neighborhoods of all network elements, i.e. a buffer of ε m around all links and nodes. Here we set $\varepsilon = 500$ m together with the grid seed distance, as this implies a theoretical coverage of 100% of the city area for a grid triangulation and an average distance to the network of 167 m, see Supplementary Note 2. In general, a cohesive bicycle network should cover the majority of the city area.

Seed point coverage. This metric refers to the number of seed points that have been covered by network elements (by the coverage defined above).

Components. The number of disconnected components is the number of maximal connected subgraphs, i.e. all pairs of nodes within one component are reachable with a path but there is no path between nodes from different components.

Directness. The directness between two nodes i and j is generally defined as the ratio $\frac{d_E(i,j)}{d_G(i,j)}$ between euclidean distance $d_E(i,j)$ and shortest path distance $d_G(i,j)$. The average of this ratio over all pairs of nodes is then the directness of the whole network:

$$D = \left\langle \frac{d_E(i,j)}{d_G(i,j)} \right\rangle_{i \neq j} \tag{3}$$

Node pairs i and j are considered from within the same components because directness is a meaningless concept for nodes from different components. Other possible definitions for directness could be:

- The previous definition but only applied to the LCC: $D = \left\langle \frac{d_E(i,j)}{d_G(i,j)} \right\rangle_{i \neq j \in LCC}$
- The ratio of total euclidian distances and shortest path distances: $D = \frac{\sum_{i \neq j} d_E(i,j)}{\sum_{i \neq j} d_G(i,j)}$
- The previous definition but only applied to the LCC: $D = \frac{\sum_{i \neq j \in LCC} d_E(i,j)}{\sum_{i \neq j \in LCC} d_G(i,j)}$

We calculated directness according to all these different definitions as a robustness check, see Supplementary Figure 3. Numerical values vary only insignificantly, all results are qualitatively identical for each definition.

Local and global efficiency. A network's global efficiency is defined as[47]:

$$E_{\text{glob}} = \frac{\sum_{i \neq j} \frac{1}{d_G(i,j)}}{\sum_{i \neq j} \frac{1}{d_E(i,j)}} \tag{4}$$

A network's local efficiency is defined as the average of global efficiencies $E_{\text{glob}}(i)$ over each node i and its neighbors,

$$E_{\text{loc}} = \frac{1}{N} \sum_{i=1}^{N} E_{\text{glob}}(i) \tag{5}$$

Local efficiency measures local fault tolerance and therefore operationalizes the concept of resilience on a local level.

Spatial clustering and anisotropy. We first specify a threshold θ and identify the N_θ nodes with high betweenness above the θ-th percentile. Then, we compute their spread about their center of mass

$$x_{\mathrm{cm}} = \frac{1}{N_\theta} \sum_{i=1}^{N_\theta} x_i$$

where x_i specifies their coordinates, normalizing for comparison across networks of different sizes via

$$C_\theta = \frac{1}{N_\theta \langle X \rangle} \sum_{i=1}^{N_\theta} \| x_i - x_{\mathrm{cm}} \|, \tag{6}$$

where

$$\langle X \rangle = \frac{1}{N} \sum_{i=1}^{N} \| x_i - x_{\mathrm{cm}} \|$$

is the average distance of all nodes in the network to the center of mass of the high betweenness cluster.

Transition between the topological and spatial regimes is quantified by the increasingly isotropic layout of the high betweenness nodes with increasing edge-density. The anisotropy factor is defined by the ratio

$$A_\theta = \frac{\lambda_1}{\lambda_2} \tag{7}$$

where $\lambda_1 \leq \lambda_2$ are the positive eigenvalues of the covariance matrix of the spatial position of the nodes with betweenness above the threshold θ[51].

For the largest 15 cities we calculated these values only at the 0, 0.5, and 1 quantiles of the growth strategies due to computational feasibility. Therefore, Supplementary Fig. 4 reports average values over the 47 smallest cities.

Data availability

All code used in the research is open-sourced, available at: https://github.com/mszell/bikenwgrowth. All data used and generated in the research are publicly available at Zenodo[83]: https://zenodo.org/record/5083049. Interactively growing networks, plots and video visualizations for all 62 cities can be explored and downloaded at the accompanying visualization platform: https://growbike.net.

Received: 8 December 2021; Accepted: 12 April 2022
Published online: 26 April 2022

References

1. Banister, D. *Unsustainable Transport: City Transport in the New Century* (Routledge, 2005).
2. Nieuwenhuijsen, M. J. & Khreis, H. Car free cities: Pathway to healthy urban living. *Environ. Int.* **94**, 251–262 (2016).
3. Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407 (2020).
4. Gössling, S., Choi, A., Dekker, K. & Metzler, D. The social cost of automobility, cycling and walking in the European Union. *Ecol. Econ.* **158**, 65–74 (2019).
5. Gössling, S. Why cities need to take road space from cars-and how this could be done. *J. Urban Des.* **1**, 1–6 (2020).
6. Szell, M. Crowdsourced quantification and visualization of urban mobility space inequality. *Urban Plan.* **3**, 1–20 (2018).
7. Creutzig, F. *et al.* Transport: A roadblock to climate change mitigation?. *Science* **350**, 911–912 (2015).
8. Milovanoff, A., Posen, I. D. & MacLean, H. L. Electrification of light-duty vehicle fleet alone will not meet mitigation targets. *Nat. Clim. Change* **10**, 1102–1107 (2020).
9. Brand, C. *et al.* The climate change mitigation effects of daily active travel in cities. *Transp. Res. D* **93**, 102764 (2021).
10. Mattioli, G., Roberts, C., Steinberger, J. K. & Brown, A. The political economy of car dependence: A systems of provision approach. *Energy Res. Soc. Sci.* **66**, 101486 (2020).
11. Feddes, F., de Lange, M. & te Brömmelstroet, M. *The Politics of Cycling Infrastructure: Spaces and (In) Equality* 133 (Policy Press, 2020).
12. Carstensen, T. A., Olafsson, A. S., Bech, N. M., Poulsen, T. S. & Zhao, C. The spatio-temporal development of Copenhagen's bicycle infrastructure 1912–2013. *Geogr. Tidsskr. Danish J. Geogr.* **115**, 142–156 (2015).
13. Natera Orozco, L. G., Battiston, F., Iñiguez, G. & Szell, M. Data-driven strategies for optimal bicycle network growth. *R. Soc. Open Sci.* **7**, 201130 (2020).
14. CROW, *Design manual for bicycle traffic* (2016).
15. Ripple, W. *et al.* World scientists' warning of a climate emergency. *BioScience* **70**(1), 8–12 (2019).
16. I. P. on Climate Change (IPCC). Climate change 2021: The physical science basis (2021).
17. Lamb, W. F. *et al.* A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018. *Environ. Res. Lett.* **17**(4), 049502 (2021).
18. Caiazzo, F., Ashok, A., Waitz, I. A., Yim, S. H. & Barrett, S. R. Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005. *Atmos. Environ.* **79**, 198–208 (2013).
19. Klanjčić, M., Gauvin, L., Tizzoni, M. & Szell, M. Identifying urban features for vulnerable road user safety in Europe. *EPJ Data Sci.* (2022).
20. Jeong, H., Ryu, J.-S. & Ra, K. Characteristics of potentially toxic elements and multi-isotope signatures (cu, zn, pb) in non-exhaust traffic emission sources. *Environ. Pollut.* **292**, 118339 (2022).
21. Pereira, R. H., Schwanen, T. & Banister, D. Distributive justice and equity in transportation. *Transp. Rev.* **37**, 170–191 (2017).
22. Lovelace, R., Morgan, M., Talbot, J. & Lucas-Smith, M. *Methods to Prioritise Pop-up Active Transport Infrastructure* (Springer, 2020).
23. Kraus, S. & Koch, N. Provisional COVID-19 infrastructure induces large, rapid increases in cycling. *Proc. Natl. Acad. Sci.* **118**, 1–10 (2021).

24. Marqués, R., Hernández-Herrador, V., Calvo-Salazar, M. & García-Cebrián, J. A. How infrastructure can promote cycling in cities: Lessons from seville. *Res. Transp. Econ.* **53**, 31–44 (2015).
25. City of Paris, Un nouveau plan vélo pour une ville 100 % cyclable (2021).
26. Zhao, C., Carstensen, T. A., Nielsen, T. A. S. & Olafsson, A. S. Bicycle-friendly infrastructure planning in Beijing and Copenhagen-between adapting design solutions and learning local planning cultures. *J. Transp. Geogr.* **68**, 149–159 (2018).
27. Boisjoly, G., Lachapelle, U. & El-Geneidy, A. Bicycle network performance: Assessing the directness of bicycle facilities through connectivity measures, a Montreal, Canada case study. *Int. J. Sustain. Transp.* **14**, 620–634 (2020).
28. Lowry, M. & Loh, T. H. Quantifying bicycle network connectivity. *Prevent. Med.* **95**, S134–S140 (2017).
29. Olmos, L. E. *et al.* A data science framework for planning the growth of bicycle infrastructures. *Transp. Res. C* **115**, 102640 (2020).
30. Palominos, N. & Smith, D. A. Identifying and characterising active travel corridors for London in response to COVID-19 using shortest path and streetspace analysis (2020).
31. Medeiros, R. M., Bojic, I. & Jammot-Paillet, Q. Spatiotemporal variation in bicycle road crashes and traffic volume in berlin: Implications for future research, planning, and network design. *Future Transp.* **1**, 686–706 (2021).
32. Mahfouz, H., Arcaute, E. & Lovelace, R. A road segment prioritization approach for cycling infrastructure. arXiv:2105.03712 (2021).
33. Batty, M. *The New Science of Cities* (MIT Press, 2013).
34. Resch, B. & Szell, M. Human-centric data science for urban studies. *ISPRS Int. J. Geo-Inf.* **8**, 584 (2019).
35. Teschke, K. *et al.* Route infrastructure and the risk of injuries to bicyclists: A case-crossover study. *Am. J. Public Health* **102**, 2336–2343 (2012).
36. Erdős, P. & Rényi, A. On random graphs. *Publ. Math.* **6**, 290–297 (1959).
37. Zeng, G. *et al.* Switch between critical percolation modes in city traffic dynamics. *Proc. Natl. Acad. Sci.* **116**, 23 (2019).
38. Gross, B., Vakhin, D., Buldyrev, S. & Havlin, S. Two transitions in spatial modular networks. *N. J. Phys.* **22**, 053002 (2020).
39. Rhoads, D., Solé-Ribalta, A., González, M. C. & Borge-Holthoefer, J. Planning for sustainable open streets in pandemic cities. arXiv:2009.12548 (2020).
40. van Nes, R. Design of multimodal transport networks, Ph.D. thesis, Civil Engineering. Delft Technical University, Delft (2002).
41. Barrington-Leigh, C. & Millard-Ball, A. The world's user-generated road map is more than 80% complete. *PLoS ONE* **12**, e0180698 (2017).
42. Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2011).
43. Boeing, G. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Comput. Environ. Urban Syst.* **65**, 126–139 (2017).
44. Zhang, X., Miller-Hooks, E. & Denny, K. Assessing the role of network topology in transportation network resilience. *J. Transp. Geogr.* **46**, 35–45 (2015).
45. Mulzer, W. & Rote, G. Minimum-weight triangulation is np-hard. *J. ACM (JACM)* **55**, 1–29 (2008).
46. Cardillo, A., Scellato, S., Latora, V. & Porta, S. Structural properties of planar graphs of urban street patterns. *Phys. Rev. E* **73**, 1–7 (2006).
47. Latora, V. & Marchiori, M. Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87**, 198701 (2001).
48. Achlioptas, D., D'Souza, R. M. & Spencer, J. Explosive percolation in random networks. *Science* **323**, 1453–1455 (2009).
49. Bollobás, B. & Thomason, A. G. Threshold functions. *Combinatorica* **7**, 35–38 (1987).
50. NACTO. *Urban Bikeway Design Guide* (Island Press, 2014).
51. Kirkley, A., Barbosa, H., Barthelemy, M. & Ghoshal, G. From the betweenness centrality in street networks to structural invariants in random planar graphs. *Nat. Commun.* **9**, 2501 (2018).
52. Schepers, P., Twisk, D., Fishman, E., Fyhri, A. & Jensen, A. The Dutch road to a high level of cycling safety. *Saf. Sci.* **92**, 264–273 (2017).
53. Rietveld, P. & Daniel, V. Determinants of bicycle use: Do municipal policies matter?. *Transp. Res. A* **38**, 531–550 (2004).
54. Schoner, J. E. & Levinson, D. M. The missing link: Bicycle infrastructure networks and ridership in 74 us cities. *Transportation* **41**, 1187–1204 (2014).
55. Ibraeva, A., de Almeida Correia, G. H., Silva, C. & Antunes, A. P. Transit-oriented development: A review of research achievements and challenges. *Transp. Res. A* **132**, 110–130 (2020).
56. Folco, P., Gauvin, L., Tizzoni, M. & Szell, M. Data-driven bicycle network planning for demand and safety. arXiv:2203.14619 (2022).
57. Vybornova, A., Cunha, T., Gühnemann, A. & Szell, M. Automated detection of missing links in bicycle networks. *Geogr. Anal.* **1**, 1–29 (2022).
58. Quercia, D., Schifanella, R. & Aiello, L. M. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 116–125 (2014).
59. Walker, J. To predict with confidence, plan for freedom. *J. Public Transp.* **21**, 12 (2018).
60. Bullard, R. D., Johnson, G. S. & Torres, A. O. *Highway Robbery: Transportation Racism & New Routes to Equity* (South End Press, 2004).
61. Hoffmann, M. L. *Bike Lanes are White Lanes: Bicycle Advocacy and Urban Planning* (University of Nebraska Press, 2016).
62. Jafino, B. A. An equity-based transport network criticality analysis. *Transp. Res. A* **144**, 204–221 (2021).
63. Nelson, A. C. & Allen, D. If you build them, commuters will use them: Association between bicycle facilities and bicycle commuting. *Transp. Res. Rec.* **1578**, 79–83 (1997).
64. Lyons, G. & Davidson, C. Guidance for transport planning and policymaking in the face of an uncertain future. *Transp. Res. A* **88**, 104–116 (2016).
65. ITF. Travel transitions: How transport planners and policy makers can respond to shifting mobility trends, *Tech. rep.* (OECD Publishing, 2021).
66. Transport strategies for net-zero systems by design, *Tech. rep.* (OECD Publishing, 2021).
67. European Commission. Reclaiming city streets for people. Chaos or quality of life? *Tech. rep.* (Directorate-General for the Environment, 2004).
68. Lovelace, R. *et al.* The propensity to cycle tool: An open source online system for sustainable transport planning. *J. Transp. Land Use* **10**, 505–528 (2017).
69. Marshall, B., De Lucia, S. & Day, H. *Transport technology tracker wave 7* (Tech. rep, UK Department for Transport, 2021).
70. Lamb, W. F. *et al.* Discourses of climate delay. *Glob. Sustain.* **3**, 1–10 (2020).
71. Cervero, R., Sarmiento, O. L., Jacoby, E., Gomez, L. F. & Neiman, A. Influences of built environments on walking and cycling: Lessons from Bogotá. *Int. J. Sustain. Transport.* **3**, 203–226 (2009).
72. Hughes, T. P. *et al.*, The evolution of large technological systems. *The social construction of technological systems: New directions in the sociology and history of technology* pp. 45–76 (2012).
73. Bijker, W. E. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change* (MIT Press, 997).
74. Nieuwenhuijsen, M. J. Urban and transport planning pathways to carbon neutral, liveable and healthy cities; a review of the current evidence. *Environ. Int.* **1**, 105661 (2020).
75. PrietoCuriel, R., GonzálezRamírez, H., QuiñonesDomínguez, M. & OrjuelaMendoza, J. P. A paradox of traffic and extra cars in a city as a collective behaviour. *R. Soc. Open Sci.* **8**(6), 201808 (2021).
76. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B* **37**, 682–703 (2010).

77. Ferster, C., Fischer, J., Manaugh, K., Nelson, T. & Winters, M. Using OpenStreetMap to inventory bicycle infrastructure: A comparison with open data from cities. *Int. J. Sustain. Transp.* **1**, 1–10 (2019).
78. Boeing, G. Urban spatial order: Street network orientation, configuration, and entropy. *Appl. Netw. Sci.* **4**, 1–19 (2019).
79. Levcopoulos, C. & Krznaric, D. Quasi-greedy triangulations approximating the minimum weight triangulation. *J. Algorithms* **27**, 303–338 (1998).
80. Toussaint, G. *Workshop on Algorithms and Data Structures* 218–227 (Springer, 1995).
81. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35 (1977).
82. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1**, 215–239 (1978).
83. Szell, M. *Urban Bicycle Networks, Existing and Synthetically Grown* (Zenodo, 2021).

Acknowledgements

We thank Cecilia Laura Kolding Andersen and Morten Lynghede for developing and implementing the visualization platform. We are grateful to Anders Hartmann, Anastassia Vybornova, and Laura Alessandretti for helpful discussions. We thank the ITU High-Performance Computing cluster for computing resources and support. We gratefully acknowledge the open data that this article is based on, from https://www.openstreetmap.org, copyright OpenStreetMap contributors.

Author contributions

M.S. designed the study with input from R.S. and G.G. M.S. wrote the manuscript with input from all authors. M.S. acquired and pre-processed the data with input from S.M. and T.P., and performed the simulations. M.S. directed the project, R.S. and G.G. helped supervise the project. M.S. measured the results, aided by S.M. T.P. performed the analysis on grid size and network coverage. All authors discussed the results.

Funding

This work was supported by the Danish Ministry of Transport.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10783-y.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Polaris: Sampling from the Multigraph Configuration Model with Prescribed Color Assortativity

Giulia Preti
giulia.preti@centai.eu
CENTAI
Turin, Italy

Matteo Riondato
mriondato@amherst.edu
Amherst College
Amherst, MA, USA

Aristides Gionis
argioni@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Gianmarco De Francisci Morales
gdfm@acm.org
CENTAI
Turin, Italy

Abstract

We introduce Polaris, a network null model for colored multigraphs that preserves the Joint Color Matrix. Polaris is specifically designed for studying network polarization, where vertices belong to a side in a debate or a partisan group, represented by a vertex color, and relations have different strengths, represented by an integer-valued edge multiplicity. The key feature of Polaris is preserving the Joint Color Matrix (JCM) of the multigraph, which specifies the number of edges connecting vertices of any two given colors. The JCM is the basic property that determines color assortativity, a fundamental aspect in studying homophily and segregation in polarized networks. By using Polaris, network scientists can test whether a phenomenon is entirely explained by the JCM of the observed network or whether other phenomena might be at play.

Technically, our null model is an extension of the configuration model: an ensemble of colored multigraphs characterized by the same degree sequence and the same JCM.To sample from this ensemble, we develop a suite of Markov Chain Monte Carlo algorithms, collectively named Polaris-$*$. It includes Polaris-B, an adaptation of a generic Metropolis-Hastings algorithm, and Polaris-C, a faster, specialized algorithm with higher acceptance probabilities. This new null model and the associated algorithms provide a more nuanced toolset for examining polarization in social networks, thus enabling statistically sound conclusions.

CCS Concepts

• **Information systems** → *Web mining*; • **Theory of computation** → *Graph algorithms analysis*; **Random walks and Markov chains**; **Generating random combinatorial structures**; **Social networks**; • **Mathematics of computing** → *Random graphs*.

Keywords

Hypothesis Testing, Null Model, Polarization

1 Introduction

Polarization is perceived as one of the largest problems in our society [18]. Scientists have studied the phenomenon extensively, more recently by using data from social media [12, 29, 32, 33, 41]. Many different theories try to explain the phenomenon, from affective polarization to partisan identity and echo chambers [3, 10, 23, 28, 35, 51]. However, definite evidence is still lacking.

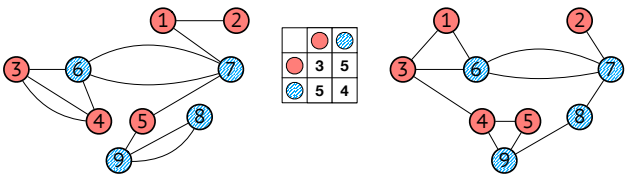


Figure 1: Two multigraphs with the same degree sequence and JCM.

Existing observational studies often use network representations to study the problem. This choice allows employing the ample network- and graph-theoretical toolset to define properties and compute relevant measures. However, such quantities are only significant insofar as they are not statistical noise. For this purpose, *null-hypothesis models* are used to assess statistical significance.

Unfortunately, to date, the network null models used in these studies are exceedingly simple [46]. They usually preserve only basic characteristics of the graph structure, such as density or degree sequences [20], but they ignore the interplay between opinions or communities that describe the polarization phenomenon.

The main contribution of this work is to propose a new network null model geared towards the study of network polarization. In particular, our null model is a *statistical ensemble of colored multigraphs*: graphs where each vertex has a color (i.e., a single label) and edges can appear multiple times. Vertex colors, or labels, are often used to represent the different sides in a controversial argument or debate, or groups such as partisan identities [12, 21, 24]. Multiedges are commonly used to represent endorsement networks (e.g., retweet or interaction networks) [21, 22, 24], where the multiplicity represents the strength of the relationship between two vertices.

The ensemble we consider is a microcanonical one akin to the configuration model [5, 6], i.e., its members are all and only the graphs with a specific degree sequence. The graph ensemble for our null model is additionally defined by a property shared by all members of the ensemble: the Joint Color Matrix (JCM). This matrix determines the number of edges that connect vertices of different colors. Figure 1 depicts two small graphs belonging to the same ensemble and their associated JCM. The JCM determines important properties of the graph, e.g., its color assortativity [40] which is fundamental in the study of homophily and segregation [34].

We devise a suite of Markov chain Monte Carlo algorithms, named Polaris-$*$, to sample from the ensemble. We prove the Markov chain is irreducible and aperiodic, thus having a unique

stationary distribution. The first algorithm, Polaris-B, is an adaptation of an existing algorithm [19] using the Metropolis-Hastings method. The second algorithm, Polaris-C, takes into account the vertex colors in a more judicious manner. As a result, Polaris-C has higher acceptance probabilities than Polaris-B, and mixes faster.

2 Related Work

When searching for patterns in network data it is essential to be able to reason about their significance. In statistics, there is a long tradition of assessing significance by comparing an observed pattern with its occurrence in a *randomized null model* [17]. Extending this idea to networks leads to *random-graph null models*, where one compares properties observed in real-world networks with properties observed in networks sampled from a certain random-graph distribution. While there is a vast literature on random-graph models, such as the Erdős-Rényi random graph [15] and the preferential attachment model [4, 7], which are simple to generate via an iterative sampling process, practitioners often seek to sample networks from a space of networks satisfying certain constraints. The most commonly-used constrained random-graph null model is the *configuration model* [5, 6, 19], where the sample space consists of all networks having a specified degree sequence. Configuration models have a long research history with applications in sociology [38], ecology [11], systems biology [36], and other disciplines.

The *Markov chain Monte Carlo (MCMC)* method is an archetypal approach for sampling from the space of networks with a fixed degree sequence. The MCMC technique performs a random walk over the sample space $\mathcal{G}$ of feasible networks and appropriately modifies the transition probabilities of the walk, e.g., using the Metropolis-Hastings algorithm [42], so that the *stationary distribution* is a desirable target distribution, such as the *uniform* one. Typically, the neighboring states in the Markov chain are networks that differ only in a *two-edge swap*, making it easy to transition between states and sample random networks from the whole space $\mathcal{G}$. By proving that the Markov chain is *strongly connected* (*irreducible*) and *aperiodic*, it can be argued that there is a *unique stationary distribution*, and the Metropolis-Hastings algorithm can be used to obtain samples from it.

A question of theoretical interest is the *mixing time* of the MCMC method, i.e., the number of steps required before the actual sample distribution is ϵ-close to the stationary distribution. Theoretical papers have derived conditions that the method is *rapidly mixing*, i.e., the number of required steps is polynomial [13, 31]. However, the general case is still not fully understood. Furthermore, existing bounds are high-degree polynomials and are mostly of theoretical interest. In practice, researchers employ various diagnostics to assess empirically the MCMC convergence [14, 45], such as comparing the variance of sample graph statistics inside a sequence of the chain and against the variance across multiple sequences [27].

Other types of graph ensembles have been proposed as a null model to assess the statistical importance of patterns and graph structure. Those include *maximum-entropy models* [48], which however preserve the degree sequence only in expectation, and *exponential random graph models* [47], which increase the probability of observing certain subgraph structures, and which are also typically sampled using MCMC methods.

Overall, a large number of methods for sampling graph null models have been presented in the literature, and the ideas have been applied to analyzing data from different disciplines. Nevertheless, perhaps surprisingly, no previous work on preserving the color assortativity of a network exists.

3 Preliminaries

This section introduces the main concepts and notation used throughout the work. We use double curly braces to denote multisets, e.g., $A = \{\!\{a, b, b, c, d, d, d\}\!\}$, and $\omega_A(a)$ is the *multiplicity* of an element a in a multiset A, i.e., the number of times a appears in A. $|A|$ denotes the multiset cardinality of a set, e.g., $|A| = 7$ in the example above. The notation $\text{set}(A)$ represents the set obtained by removing all duplicates from A, e.g., $\text{set}(A) = \{a, b, c, d\}$.

Definition 3.1 (Colored Multigraph). A *colored, undirected, multigraph* is a tuple $G \doteq (V, E, \mathcal{L}, \lambda)$, where V is a set of vertices, E is a multiset of edges between vertices, each edge being an unordered pair of vertices from V, and $\lambda : V \to \mathcal{L}$ is a labeling function assigning a color (i.e., a single label) from the set $\mathcal{L}$ to each vertex.

We allow (multiple) self-loops, i.e., edges of the type $(u, u), u \in V$. All multigraphs we consider are colored, so henceforth we use "multigraph" to mean "colored multigraph". We refer to edges incident to vertices with the same color as *monochrome* edges, and edges incident to vertices with different colors as *bichrome* edges.

Two edges $(u, w), (v, z)$ are *distinct* when they are two different members of E. Two distinct edges may be *copies* of the same multiedge, i.e., be incident to the same pair of vertices, or to the same vertex if they are self-loop. We write $(u, w) = (v, z)$ if two edges are copies, but since edges are unordered pairs of vertices, this notation does not imply $u = v$ and $w = z$, as it may be $u = z$ and $w = v$.

A multigraph $G = (V, E, \mathcal{L}, \lambda)$ can be seen as an integer-weighted graph $G' = (V, E', \mathcal{L}, \lambda, \mathrm{w})$, with $E' = \text{set}(E)$, and w a function that assigns a natural weight to edges in E', so that $\mathrm{w}(e) = \omega_E(e)$.

Given a multigraph $G = (V, E, \mathcal{L}, \lambda)$, for each $u \in V$, $\Gamma_G(u)$ denotes the multiset of neighbors of u, and $d_G(u) \doteq |\Gamma_G(u)|$ the *degree of u in G*. For each $\ell \in \mathcal{L}$, let $V^\ell \doteq \{v \in V : \lambda(v) = \ell\}$ be the set of vertices with color ℓ. For each $u \in V$ and $\ell \in \mathcal{L}$, let $\Gamma_G^\ell(u) \doteq \{\!\{v \in V^\ell : (u, v) \in E\}\!\}$ be the multiset of neighbors of u in G with color ℓ, and let $\gamma_G^\ell(u) \doteq \left|\Gamma_G^\ell(u)\right|$. Clearly, $d_G(u) = \sum_{\ell \in \mathcal{L}} \gamma_G^\ell(u)$.

Definition 3.2 (JCM). The *Joint Color Matrix* J_G of a multigraph $G = (V, E, \mathcal{L}, \lambda)$ is the symmetric square matrix $\mathrm{J}_G \in \mathbb{N}^{|\mathcal{L}| \times |\mathcal{L}|}$ where each entry $J_G[\ell, r]$ is the number of edges between a vertex with color ℓ and a vertex with color r, i.e.,

$$J_G[\ell, r] \doteq |\{\!\{(u, w) \in E : \lambda(u) = \ell \wedge \lambda(w) = r\}\!\}| \ .$$

Figure 1 shows two multigraphs with two colors, the same degree sequence, and the same JCM (shown in the center of the figure).

3.1 Null Models for Graph Properties

For any multigraph G, let $\mathcal{P}_G$ be a set of properties from G, e.g., the number of edges, the degree sequence, the diameter, or similar structural properties, which may be scalars, vectors, or matrices.

Let $\mathring{G} = (V, E, \mathcal{L}, \lambda)$ be an observed multigraph. Given $\mathcal{P}_{\mathring{G}}$, the microcanonical *null model* $\Pi \doteq (\mathcal{Z}, \pi)$ is a tuple where $\mathcal{Z}$ is the set of all and only the multigraphs $G = (V, E_G, \mathcal{L}, \lambda)$ on the same set

of vertices, with the same colors and coloring function as $\mathring{G}$, and that preserve each property in $\mathcal{P}_{\mathring{G}}$ i.e., such that $\mathcal{P}_G = \mathcal{P}_{\mathring{G}}$, and where π is a probability distribution over $\mathcal{Z}$. Clearly $\mathring{G} \in \mathcal{Z}$, and the graphs in $\mathcal{Z}$ only differ by their multisets of edges.

3.2 Markov Chain Monte Carlo Methods

All algorithms in Polaris-$*$ follow the *Markov chain Monte Carlo (MCMC) method*, using the *Metropolis-Hastings (MH) approach* [37, Ch. 7 and 10]. Let us now introduce these concepts.

Let $\mathcal{G} = (\mathcal{S}, \mathcal{E}, \mathrm{w})$ be a directed, weighted, strongly connected, and aperiodic[1] graph, which may have self-loops. The vertices in $\mathcal{S}$ are known as *states*, and $\mathcal{G}$ is known as a state graph. Using the same notation we defined previously, for any state $s \in \mathcal{S}$, $\Gamma_{\mathcal{G}}(s)$ denotes the set of (out-)*neighbors* of s, i.e., the set of states u s.t. $(s,u) \in \mathcal{E}$. Iff $u \in \Gamma_{\mathcal{G}}(s)$, then $\mathrm{w}(s,u) > 0$, and it holds $\sum_{u \in \Gamma_{\mathcal{G}}(s)} \mathrm{w}(s,u) = 1$. Thus, for any $u \in \mathcal{S}$, we can define the *transition probability* $\tau_{s,u}$ *from* s *to* u as $\mathrm{w}(s,u)$ if $u \in \Gamma_{\mathcal{G}}(s)$, and 0 otherwise.

Given any $\mathcal{G}$ as above, a *neighbor proposal probability distribution* ξ_v over $\Gamma_{\mathcal{G}}(v)$ for any $v \in \mathcal{S}$, and any probability distribution ϕ over $\mathcal{S}$, the *MH approach* is a generic procedure to sample a state $s \in \mathcal{S}$ according to ϕ. Starting from any $v \in \mathcal{S}$, one first draws a neighbor u of v according to ξ_v, and then "moves" to u with probability

$$\alpha_v(u) = \min\left\{1, \frac{\phi(u)}{\phi(v)} \frac{\xi_u(v)}{\xi_v(u)}\right\},$$

otherwise remains in v. The quantity $\alpha_v(u)$ is known as the *acceptance probability* of u. The sequence of states obtained by repeating this procedure forms a Markov chain over $\mathcal{S}$ with unique stationary distribution ϕ. Thus, after a sufficiently large number of steps t, the state v_t at time t is distributed according to ϕ, and can be considered a sample of $\mathcal{S}$ according to ϕ.

To use MH, it is necessary to define: *(i)* the graph $\mathcal{G}$ as above, taking special care in ensuring that it is strongly-connected and aperiodic; *(ii)*, the neighbor sampling distribution $\xi_s()$ for every state $s \in \mathcal{S}$; and *(iii)* the desired sampling distribution ϕ over $\mathcal{S}$.

4 A Null Model for Vertex-Colored Graphs

Given an observed $\mathring{G} \doteq (V, E, \mathcal{L}, \lambda)$, with $V = \{v_1, \ldots, v_{|V|}\}$, we consider the null model $\Pi = (\mathcal{Z}, \pi)$ where $\mathcal{P}_{\mathring{G}}$ consists of the degree sequence $\left[d_{\mathring{G}}(v_1), \ldots, d_{\mathring{G}}\left(v_{|V|}\right)\right]$ and the JCM $J_{\mathring{G}}$.

This null model is essentially the simplest one that considers the color information, if one assumes that the color of a vertex is an intrinsic property. While one could think of preserving only the colored degree sequences for each color, doing so is equivalent to preserving the "generic" degree sequence, and thus does not leverage the color information in any meaningful way. Indeed, this can be done on the unlabeled version of the multigraph [20].

Our goal is to design efficient MCMC algorithms to sample from $\mathcal{Z}$ w.r.t. π as defined above. We first define two operations that allow transforming a multigraph G into a multigraph H, potentially identical to G. The first operation is the classic Double Edge Swap (DES), known under many names and introduced many times in the literature [2, 8, 30, 49, 50, 52].

[1] A graph is aperiodic iff the greatest common divisor of the lengths of its cycles is 1.

Definition 4.1 (Double Edge Swap (DES)). Given a multigraph $G \doteq (V, E, \mathcal{L}, \lambda)$, let $(u,w), (v,z)$ be two distinct edges in E. Consider the multigraph $H = (V, (E \setminus \{\!\{(u,w),(v,z)\}\!\}) \cup \{\!\{(u,z),(w,v)\}\!\}, \mathcal{L}, \lambda)$. We call the operation that "swaps" $(u,w), (v,z)$ with $(u,z), (w,v)$ a *Double Edge Swap (DES)*, and denote it $(u,w), (v,z) \to (u,z), (w,v)$.

We say that a DES is *applied to* the origin multigraph G to obtain the destination multigraph H, or that a DES *transforms* G *into* H.

For every unordered pair $((u,w),(v,z))$ of distinct edges in the origin graph, there are exactly two DESs that involve them: $(u,w), (v,z) \to (u,z), (v,w)$ and $(u,w), (v,z) \to (u,v), (z,w)$. If the destination multigraph H is the same for both DESs, we say that the DESs are *equivalent*. If $H = G$, we say that the DES is a *no-op*, otherwise we say that the DES is a *moving* DES. For the same unordered pair of distinct edges in the origin graph, one DES may be a no-op, and the other may be a moving DES.

Multiple expressions may correspond to the same DES, as a DES is defined by the multiset of edges in the origin multigraph and by the multiset of edges in the destination multigraph. For example, the expressions $(u,w), (v,z) \to (u,z), (w,v)$ and $(z,v), (u,w) \to (u,z), (w,v)$ both denote the same DES.

DESs can be used in MCMC algorithms to sample from a null model that preserves the degree sequence of an observed multigraph [20, and references therein]: given a DES, the destination multigraph has the same vertices and the same degree sequence as the origin. Conversely, the JCM may or may not be preserved by a DES. Thus we define the following specific operation.

Definition 4.2 (JCM-preserving Double Edge Swap (JDES)). A *JCM-preserving Double Edge Swap (JDES)* is a DES such that the destination multigraph H retains the JCM of the origin multigraph G.

An example of JDES that can be applied to the left multigraph in Figure 1 is $(1,7), (3,6) \to (1,6), (3,7)$, while the operation $(3,4), (9,8) \to (3,8), (9,4)$ is a DES but not a JDES.

For any unordered pair $((u,w),(v,z))$ of distinct edges in the origin multigraph, zero, one, or both DESs may be JDESs. We now give a complete characterization of which DESs are JDESs, considering different cases based on the properties of the edges involved.

Case 0: $\{\lambda(u), \lambda(w)\} \cap \{\lambda(v), \lambda(z)\} = \emptyset$, i.e., the two edges have disjoint vertex colors. Then, neither DES is a JDES.

Case 1: $|\{u,w,v,z\}| = 1$, i.e., (u,w) and (v,z) are copies of the same self-loop multiedge. Both DESs are JDESs and no-ops.

Case 2A: $|\{u,w,v,z\}| = 2 \wedge u = w \wedge v = z \wedge \lambda(u) = \lambda(v)$, i.e., (u,w) and (v,z) are two different self-loops on vertices with the same color. Both DESs are equivalent JDESs and moving DESs.

Case 2B: $|\{u,w,v,z\}| = 2 \wedge u \neq w \wedge v \neq z \wedge \lambda(u) = \lambda(w)$, i.e., (u,w) and (v,z) are identical non-self-loop monochrome multiedges. Both DESs are JDESs; one is a no-op, while the other creates a self-loop.

Case 2C: $|\{u,w,v,z\}| = 2 \wedge u \neq w \wedge v \neq z \wedge \lambda(u) \neq \lambda(w)$, i.e., (u,w) and (v,z) are identical non-self-loop bichrome multiedges. Only one DES is a JDES, and is a no-op.

Case 2D: $|\{u,w,v,z\}| = 2 \wedge (u = w \vee v = z) \wedge \neg(u = w \wedge v = z)$, i.e., one edge is a self-loop, the other is not but is incident to the self-loop vertex. Both DESs are JDESs and no-ops.

Case 3A: $|\{u,w,v,z\}| = 3 \wedge (w = u \vee v = z) \wedge \{\lambda(u), \lambda(w)\} \cap \{\lambda(v), \lambda(z)\} \neq \emptyset$, i.e., one edge is a self-loop, the other is not and is

incident to different vertices than the self-loop, and at least one of these vertices has the same color as the self-loop. Both DESs are equivalent JDESs and moving DESs.

Case 3B: $|\{u,w,v,z\}| = 3 \wedge u \neq w \wedge v \neq z \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 1$, i.e., neither edge is a self loop, and since $|\{u,w,v,z\}| = 3$, it holds exactly one of $u = v$, $u = z$, $w = v$, or $w = z$, so the edges form a wedge with vertices sharing the same color. Assume, w.l.o.g., that $u = v$. Then both DESs are JDESs; one is a no-op, while the other creates a self-loop on u and an edge between the vertices at the extremes of the former wedge.

Case 3C: $|\{u,w,v,z\}| = 3 \wedge u \neq w \wedge v \neq z \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 2 \wedge (\lambda(u) = \lambda(w) \vee \lambda(v) = \lambda(z))$. Similar to Case 3B, but exactly one edge is monochrome. Both DESs are JDESs; one is a no-op, while the other creates a self-loop on u and an edge between the two extremes of the former wedge.

Case 3D: $|\{u,w,v,z\}| = 3 \wedge u \neq w \wedge v \neq z \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 2 \wedge \lambda(u) \neq \lambda(w) \wedge \lambda(v) \neq \lambda(z)$. The edges form a wedge where the endpoints of the wedge have the same color and the vertex in the middle has a different color. One DES is a JDES and is a no-op.

Case 3E: $|\{u,w,v,z\}| = 3 \wedge u \neq w \wedge v \neq z \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 3$. The edges form a wedge, but all three vertices have different colors. One DES is a JDES and is a no-op.

Case 4A: $|\{u,w,v,z\}| = 4 \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 3 \wedge \lambda(u) \neq \lambda(w) \wedge \lambda(v) \neq \lambda(z)$. The edges are incident to four distinct vertices, neither of them is monochrome, they are not both bichrome with the same two colors, but they are incident to one vertex with the same color. One DES is a JDES and is a moving DES.

Case 4B: $|\{u,w,v,z\}| = 4 \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| = 2 \wedge \lambda(u) \neq \lambda(w) \wedge \lambda(v) \neq \lambda(z)$. The edges are incident to four distinct vertices and are both bichrome with the same two colors. One DES is a JDES and is a moving DES.

Case 4C: $|\{u,w,v,z\}| = 4 \wedge |\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| \in \{1,2\} \wedge (\lambda(u) = \lambda(w) \vee \lambda(v) = \lambda(z))$. The edges are incident to four distinct vertices, with at least three of them having the same color. Both DESs are JDESs, they are not equivalent, and both are moving DESs.

There cannot be more than two JDESs transforming G into H, for $H \neq G$. If there are two, they involve the same pair of edges and are equivalent. All JDESs are reversible: if there are JDESs transforming G into H, then there are JDESs transforming H into G.

4.1 Strong Connectivity and Aperiodicity of $\mathcal{Z}$ via JDESs

In Polaris, the state space $\mathcal{S}$ of the state graph $\mathcal{G} = (\mathcal{S}, \mathcal{E}, \mathrm{w})$ is $\mathcal{Z}$, and the desired probability distribution according to which to sample is π. The edge set $\mathcal{E}$ is defined as follows. For any $G, H \in \mathcal{Z}$, there is an edge $(G,H) \in \mathcal{E}$ if there is a JDES from $G \in \mathcal{Z}$ to $H \in \mathcal{Z}$. Clearly, if that is the case, there is also an edge $(H,G) \in \mathcal{E}$ as all JDESs are reversible. Additionally, there may be a self-loop in G even if there is no JDES from G to G, but G has a neighbor H such that the acceptance probability $\alpha_G(H)$ is strictly less than 1.

We say that G and H are *neighbors* iff there is a JDES transforming G into H. As required by MH, we show that the resulting state graph is strongly connected (Theorem 4.3) and aperiodic (Theorem 4.4) (full proofs in Appendix A.1), which ensures the chain is ergodic.

Theorem 4.3. *The state space $\mathcal{Z}$ is strongly connected by JDESs.*

The proof of this theorem explicitly builds a sequence of JDESs from any state $G \in \mathcal{Z}$ to any other $H \in \mathcal{Z}$, by first going from G to a $\tilde{G} \in \mathcal{Z}$ such that every vertex has in $\tilde{G}$ exactly the same number of neighbors of each color as it has in H, then going from $\tilde{G}$ to H.

Theorem 4.4. *Given a multigraph G, if either of the following conditions holds, then the state graph $\mathcal{G}$ is aperiodic:*

- *there exist two edges (u,w) and (v,z) that fall in cases 1, 2A, 2B, 2C, 2D, 3A, 3B, 3C, 3D, 3E, or 4C of the classification; or*
- *there exist a color $\ell \in \mathcal{L}$ such that there are bichrome edges (u,v), (u,z), (w,x), with all of u, v, z, w and x distinct, and $\lambda(u) = \lambda(w) = \ell$.*

The proof of Theorem 4.4 involves a case-by-case analysis of the JDES, showing that either there must be a self-loop on a vertex in $\mathcal{G}$, or there are cycles of length 3 and 2, thus ensuring aperiodicity.

The conditions in Theorem 4.4 are extremely mild. For example, the first condition implies that if there is a color such that there are two monochrome edges with that color, then the state graph is aperiodic. Only a (relatively) small class of unusual multigraphs results in periodic state graphs. Additionally, the conditions are not necessary for the graph to be aperiodic: the algorithms we present run Markov chains on a state graph that may have additional self-loops, as they are based on MH.

4.2 A first baseline algorithm

To warm up, we present a baseline algorithm Polaris-B, which is an adaptation of Fosdick et al. [19, Algorithm 3][2] to our task of interest. Fosdick et al. [19] introduced the algorithm to sample uniformly from the space of unlabeled multigraphs with the same degree sequence. Polaris-B is a tailored version of this algorithm to sample according to any distribution π from the space of colored multigraphs with the same degree sequence and the same JCM.

Polaris-B (pseudocode in Algorithm 1) starts by setting the current state G of the Markov chain to the observed multigraph $\mathring{G}$ (Line 1). It then enters a loop for t iterations. At each iteration, it first samples two edges e_1 and e_2 uniformly at random from the population of ordered pairs of distinct edges (Lines 4–5). The algorithm then randomly chooses one of the two possible DESs involving e_1 and e_2 (Line 6). If the selected DES des is not a JDES (Line 7), the algorithm samples a new DES; if it is a no-op (Line 8), the Markov chain stays in G. Otherwise, the algorithm computes a value ρ that depends on properties of the sampled edges, which is used as follows to ensure that the stationary distribution of the Markov chain is π. Let $H \neq G$ be the multigraph obtained by applying the JDES 'des' to G (Line 21). Polaris-B checks if $\rho\pi(H)/\pi(G)$ is greater than a real number sampled uniformly at random from $[0,1]$, and if so, it sets the current state G of the Markov chain to H (Line 22), otherwise the chain remains in G.

The only difference in Polaris-B w.r.t. [19, Algorithm 3] is that it checks if the sampled DES is a JDES, and keeps sampling a new DES until it is a JDES (Line 7).

Theorem 4.5. *The Markov chain run by Polaris-B has stationary distribution π.*

[2]This algorithm only appears in the arXiv version of this paper [20].

Algorithm 1: Polaris-B

Input: Observed multigraph $\mathring{G} \doteq (V, E, \mathcal{L}, \lambda)$, distribution π over $\mathcal{Z}$, number of iterations t
Output: Multigraph drawn from $\mathcal{Z}$ according to π

1 $G \leftarrow \mathring{G}$
2 **repeat** t **times**
3 **do**
4 $e_1 = (u, w) \leftarrow$ edge drawn u.a.r. from E
5 $e_2 = (v, z) \leftarrow$ edge drawn u.a.r. from $E \setminus \{e_1\}$
6 des $= (e_1, e_2 \to e'_1, e'_2) \leftarrow$ DES drawn u.a.r. from $\{(u, w), (v, z) \to (u, z), (v, w),\ (u, w), (v, z) \to (u, v), (w, z)\}$
7 **while** des *is not a JDES* // Case 0
8 **if** des *is a no-op* **then continue** // Cases 1, 2C, 2D, 3D, 3E, and no-ops for other cases
9 **if** $|\{u, w, v, z\}| = 4$ **then** // Cases 4(A,B,C), moving DES
10 $\rho \leftarrow {(\omega_G(e'_1)+1)(\omega_G(e'_2)+1)}\big/{\omega_G(e_1)\omega_G(e_2)}$
11 **else if** $|\{u, w, v, z\}| = 3$ **then**
12 **if** e_1 *is a self-loop* **or** e_2 *is a self-loop* **then** // Case 3A
13 $\rho \leftarrow {(\omega_G(e'_1)+1)(\omega_G(e'_2)+1)}\big/{2\omega_G(e_1)\omega_G(e_2)}$
14 **else** // Cases 3B and 3C, moving DES
15 $\rho \leftarrow {2(\omega_G(e'_1)+1)(\omega_G(e'_2)+1)}\big/{\omega_G(e_1)\omega_G(e_2)}$
16 **else** // i.e., $|\{u, w, v, z\}| = 2$
17 **if** *both* e_1 *and* e_2 *are self-loops* **then** // Case 2A
18 $\rho \leftarrow {(\omega_G(e'_1)+2)(\omega_G(e'_1)+1)}\big/{4\omega_G(e_1)\omega_G(e_2)}$
19 **else** // Case 2B
20 $\rho \leftarrow {4(\omega_G(e'_1)+1)(\omega_G(e'_2)+1)}\big/{\omega_G(e_1)(\omega_G(e_1)-1)}$
21 $H \leftarrow$ apply des to G
22 **if** $\mathsf{Uniform}(0,1) < {\rho\pi(H)}\big/{\pi(G)}$ **then** $G \leftarrow H$
23 **return** G

The complete proof is in Appendix A.1, and shows that Polaris-B follows the MH approach, thus ensuring the thesis.

In practice, the number of steps t must be chosen in such a way that the multigraph returned by Polaris-B is, at least approximately, distributed according to π, i.e., t should be greater or equal to the mixing time for the Markov chain. Theoretical results for the mixing time of these Markov chains are hard to obtain: even in the case of the state space of multigraphs connected by DESs (i.e., when only the degree sequence is preserved), upper bounds on the mixing time are only known in limited cases [16]. Therefore, Section 5 presents an empirical evaluation of the behavior of t.

4.3 An algorithm tailored to the task

We now present Polaris-C, a color-aware algorithm to sample a multigraph from $\mathcal{Z}$ according to π by leveraging the properties of the data and of the task better than Polaris-B (Section 4.2). As the results of our experimental evaluation show (Section 5), this algorithm has higher acceptance probability and converges faster than the baseline presented earlier.

Polaris-C improves over Polaris-B in several ways:

- it avoids sampling pairs of distinct edges falling in Case 0 of the characterization of JDES, i.e., pairs of distinct edges such that neither DES involving them is a JDES;
- if the sampled pair of distinct edges is such that one of the DESs is a no-op or not a JDES and the other is a moving JDES (Cases 2B, 2C, 3B, 3C) Polaris-C always chooses the moving one;
- if the sampled pair of distinct edges is such that the JDESs involving them are equivalent (Cases 1, 2A, 2D, 3A), it deterministically chooses one thus avoiding random choices.

As Polaris-C avoids selecting no-op JDESs in some cases (second bullet above), one should ask whether the resulting state graph is still aperiodic under the conditions stated in Theorem 4.4. The answer is that the first condition should be modified to hold only for Cases 1, 2C, 2D, 3D, and 4C. The condition for 4C is particularly mild: it only requires the existence of a color $\ell \in \mathcal{L}$ such that there are two non-self-loop, non-copies, monochrome edges with color ℓ, i.e., two edges involving four distinct vertices with the same color.

All the above improvements of Polaris-C over Polaris-B reduce the probability that the Markov chain remains in the current state, while not decreasing (and potentially increasing) the transition probability from a state to any of its different neighbors. In other words, the off-diagonal entries of the transition matrix of the Markov chain realized by Polaris-C are not smaller than the corresponding entries in the transition matrix of the Markov chain realized by Polaris-B. Polaris-C therefore precedes Polaris-B in Peskun's order [42], which implies that it has a smaller mixing time, i.e., requires fewer steps for the state of the chain to be (approximately) distributed according to the stationary distribution.

Polaris-C takes into account the number of *different* colors $\mathsf{nl}(A)$ in a multiset of vertices A to distinguish various cases of the JDES. Formally, w.l.o.g, let $\mathcal{L} = \{0, \ldots, k-1\}$, for some $k > 1$. For any unordered $(\ell, r) \in \mathcal{L} \times \mathcal{L}$, let $E_{G,\ell,r}$ be the multiset of edges incident to one vertex with color ℓ and one vertex with color r. Clearly $E_{G,\ell,\ell}$ is the multiset of the monochrome edges with color ℓ. We also define

$$E_{G,\ell} \doteq \bigcup_{r \in \mathcal{L}} E_{G,\ell,r}$$

as the multiset of edges incident to at least one vertex with color ℓ. Given a multiset A of vertices, let

$$\mathsf{nl}(A) \doteq |\{\lambda(v) : v \in A\}|.$$

Algorithm 2 presents Polaris-C's pseudocode. The algorithm takes as input the observed multigraph $\mathring{G}$, the distribution π according to which one wants to sample from $\mathcal{Z}$, and a number t of iterations. It keeps track of the current state of a Markov chain on $\mathcal{Z}$ in a variable G initialized to $\mathring{G}$ (Line 1). Polaris-C then enters a loop for t iterations. At each iteration, it first samples a color ℓ from $\mathcal{L}$ uniformly at random (Line 3), then it draws two distinct edges (u, w) and (v, z) uniformly at random respectively from $E_{G,\ell}$ and from $E_{G,\ell} \setminus \{(u, w)\}$ (Lines 4–5). It then checks which case of the JDES characterization should be considered. It either sets the variable jdes to be a moving JDES involving (u, w) and (v, z), possibly by choosing it uniformly at random when there are two non-equivalent, moving, JDESs (only happens in Case 4C, Lines 39–43), or keeps the state of the Markov chain to be the current multigraph G if the JDESs in the considered case are both no-ops (Cases 2C, 2D, 3D). In the cases when jdes is set, the algorithm also sets the variable ρ to a value that, as we discuss in the analysis of Polaris-C (Theorem 4.6), ensures that the multigraph returned by the algorithm is drawn from $\mathcal{Z}$ according to π. Let now H be the multigraph obtained by applying jdes to G. Polaris-C checks whether a value chosen uniformly at random in $[0, 1]$ is smaller than $\rho\pi(H)/\pi(G)$, and if so, updates the state G of the Markov chain to H (Line 46), otherwise the chain remains in the current state. After t iterations, the current state G is returned.

Theorem 4.6. *The Markov chain run by Polaris-C has stationary distribution π.*

The proof can be found in Appendix A.1. It essentially shows that, no matter into what case of the classification the sampled JDES falls, the algorithm follows the MH approach for choosing the acceptance probability to ensure the thesis of the theorem.

5 Experimental evaluation

Our experimental evaluation has three objectives. First, we demonstrate the qualitative differences between multigraphs sampled using the traditional configuration model and those obtained from Polaris. We focus on the configuration model as it is the standard reference model in network analysis [39] and aligns with the focus of this paper on microcanonical ensembles. Second, we analyze the extent to which the baseline algorithm Polaris-B differs from the color-aware algorithm Polaris-C in their respective movements within the state space. Lastly, we show the scalability of both Polaris-B and Polaris-C, particularly in relation to the number of vertex colors and the number of edges.

Datasets. We consider 11 real-world labeled networks, whose characteristics are summarized in Table 1 in Appendix A.2.

Experimental Setup. All experiments are run on an Intel Xeon Silver 4210R CPU@2.40GHz running FreeBSD with 383 GiB of RAM. We evaluate three sampling algorithms: the baseline color-agnostic algorithm Polaris-B, the color-aware algorithm Polaris-C, and the traditional configuration model (CM), which samples from the state space of multigraphs with a prescribed degree sequence. The code and the datasets used are available on GitHub.[3]

For the experiments aimed at the first goal, we allow 10 Markov chains to evolve for $4000m$ iterations, where m is the number of multiedges, recording the degree assortativity of the current state every $0.05m$ iterations. For the experiments aimed at the other two goals, we generate 100 independent samples by using each sampler for $m\log(m)$ iterations.

5.1 Comparison with the Configuration Model

Figure 2 shows that the color assortativity values of the multigraphs sampled by CM significantly diverge from those of the corresponding observed multigraphs, with relative errors close to 1. This discrepancy arises because CM disrupts the original correlations in the observed datasets, generating random graphs with low color assortativity. The effect is more pronounced in datasets with a larger number of colors or higher color assortativity, where the gap between the observed assortativity and that of the randomized graphs is larger. Consequently, we observe larger relative errors in datasets such as Trivago ($|\mathcal{L}| = 160$), Obamacare, and Abortion (assortativity 0.95), and smaller errors in datasets such as Comb and Guns ($|\mathcal{L}| = 2$ with assortativity values of 0.31 and 0.35, respectively). This result proves that CM does not adequately capture the color assortativity present in the observed data.

Figure 3 presents the running time of each sampler across the different datasets. This plot highlights that, despite Polaris-B and Polaris-C performing more complex operations and needing to update more quantities after each swap operation, their running

[3] https://github.com/lady-bluecopper/Polaris

Algorithm 2: Polaris-C

Input: Observed multigraph $\mathring{G} \doteq (V, E, \mathcal{L}, \lambda)$, distribution π over $\mathcal{Z}$, number of iterations t
Output: Multigraph drawn from $\mathcal{Z}$ according to π

$G \leftarrow \mathring{G}$
repeat t **times**
  $\ell \leftarrow$ color drawn u.a.r. from $\mathcal{L}$
  $(u, w) \leftarrow$ edge drawn u.a.r. from $E_{G,\ell}$
  $(v, z) \leftarrow$ edge drawn u.a.r. from $E_{G,\ell} \setminus \{(u, w)\}$
  if $|\{u, w, v, z\}| = 1$ **then continue** // Case 1
  else if $|\{u, w, v, z\}| = 2$ **then**
    if *both* (u, w) **and** (v, z) *are self-loops* **then** // Case 2A
      jdes $\leftarrow (u, u), (v, v) \rightarrow (u, v), (v, u)$
      $\rho \leftarrow (\omega_G((u, v)) + 2)(\omega_G((u, v)) + 1) \big/ \omega_G((u, u))\omega_G((v, v))$
    else if *neither* (u, w) *nor* (v, z) *is a self-loop* **and** $\lambda(u) = \lambda(w)$ **then** // Case 2B
      W.l.o.g. let $u = z$ (thus $w = v$)
      jdes $\leftarrow (u, v), (v, u) \rightarrow (u, u), (v, v)$
      $\rho \leftarrow (\omega_G((u, u)) + 1)(\omega_G((v, v)) + 1) \big/ \omega_G((u, v))(\omega_G((u, v)) - 1)$
    else continue // Case 2C or 2D
  else if $|\{u, w, v, z\}| = 3$ **then**
    if *either* (u, w) *or* (v, z) *is a self-loop* **then** // Case 3A
      W.l.o.g. let (u, w) be the self-loop
      jdes $\leftarrow (u, u), (z, v) \rightarrow (u, v), (z, u)$
      $\rho \leftarrow (\omega_G((u, v)) + 1)(\omega_G((u, z)) + 1) \big/ \omega_G((u, u))\omega_G((v, z))$
    else // W.l.o.g. assume $u = v$
      if $\mathsf{nl}(u, w, v, z) = 1$ **then** // Case 3B
        jdes $\leftarrow (u, w), (z, u) \rightarrow (u, u), (z, w)$
        $\rho \leftarrow (\omega_G((u, u)) + 1)(\omega_G((w, z)) + 1) \big/ \omega_G((u, w))\omega_G((u, z))$
      else if $\lambda(u) = \lambda(w)$ **or** $\lambda(v) = \lambda(z)$ **then** // Case 3C
        jdes $\leftarrow (u, w), (z, u) \rightarrow (u, u), (z, w)$
        $\rho \leftarrow (\omega_G((u, u)) + 1)(\omega_G((w, z)) + 1) \big/ \omega_G((u, w))\omega_G((u, z))$
      else continue // Case 3D or 3E
  else // i.e., $|\{u, w, v, z\}| = 4$
    if $\mathsf{nl}(u, w, v, z) = 3$ **and** $\lambda(u) \neq \lambda(w)$ **and** $\lambda(v) \neq \lambda(z)$ **then** // Case 4A
      W.l.o.g. let $\lambda(u) = \lambda(v)$
      jdes $\leftarrow (u, w), (v, z) \rightarrow (u, z), (v, w)$
      $\rho \leftarrow (\omega_G((u, z)) + 1)(\omega_G((v, w)) + 1) \big/ \omega_G((u, w))\omega_G((v, z))$
    else if $\mathsf{nl}(u, w, v, z) = 2$ **and** $\lambda(u) \neq \lambda(w)$ **and** $\lambda(v) \neq \lambda(z)$ **then** // Case 4B
      W.l.o.g. assume $\ell = \lambda(u) = \lambda(v)$ and let $\ell' \doteq \lambda(w) = \lambda(z)$ (it holds $\ell \neq \ell'$)
      jdes $\leftarrow (u, w), (v, z) \rightarrow (u, z), (v, w)$
      $\rho \leftarrow \dfrac{\frac{(\omega_G((u,z))+1)(\omega_G((v,w))+1)}{|E_{G,\ell}|(|E_{G,\ell}|-1)} + \frac{(\omega_G((u,z))+1)(\omega_G((v,w))+1)}{|E_{G,\ell'}|(|E_{G,\ell'}|-1)}}{\frac{\omega_G((u,w))\omega_G((v,z))}{|E_{G,\ell}|(|E_{G,\ell}|-1)} + \frac{\omega_G((u,w))\omega_G((v,z))}{|E_{G,\ell'}|(|E_{G,\ell'}|-1)}}$
    else // Case 4C
      if `fairCoinFlip()` *is head* **then**
        jdes $\leftarrow (u, w), (v, z) \rightarrow (u, z), (v, w)$
        $\rho \leftarrow (\omega_G((u, z)) + 1)(\omega_G((v, w)) + 1) \big/ \omega_G((u, w))\omega_G((v, z))$
      else
        jdes $\leftarrow (u, w), (z, v) \rightarrow (u, v), (z, w)$
        $\rho \leftarrow (\omega_G((u, v)) + 1)(\omega_G((z, w)) + 1) \big/ \omega_G((u, w))\omega_G((v, z))$
  $H \leftarrow$ apply jdes to G
  if $\mathsf{Uniform}(0, 1) < \rho\pi(H) \big/ \pi(G)$ **then** $G \leftarrow H$
return G

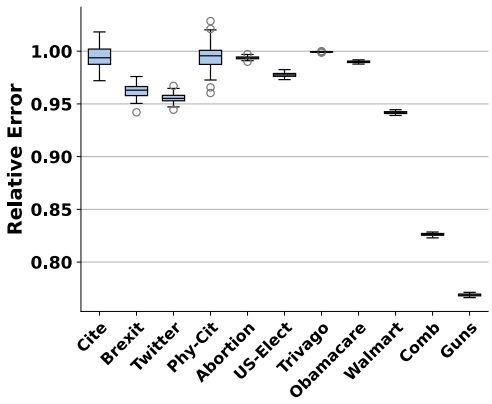


Figure 2: Distribution of the relative errors of color assortativity for samples generated by CM, compared to the color assortativity of the observed datasets, for datasets of increasing size. Results are based on 100 samples. Bars indicate one standard deviation.

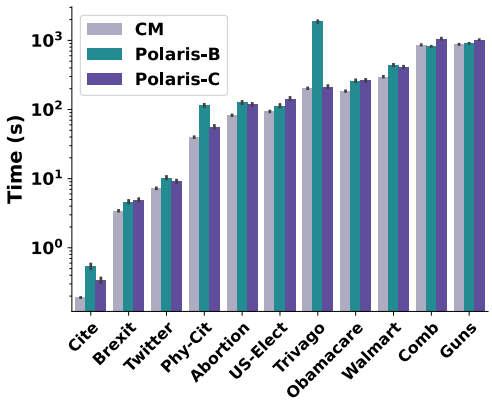


Figure 3: Running time required by each sampler to perform $m\log(m)$ iterations, for datasets of increasing size. Results for 100 samples. Bars indicate one standard deviation.

time is similar to that of CM. However, the differences in performance become especially visible in datasets with a larger number of labels, such as Trivago and Phy-Cit. In these datasets, Polaris-B takes, on average, one order of magnitude longer to generate a sample compared to the other two samplers. As the number of colors increases, the likelihood that the sampled DES is not a JDES also increases, thus increasing the running time. As a consequence, the algorithm must repeatedly sample new DESs until it finds one that is a JDES, which adds considerable overhead to the process.

Figure 4 presents the running time required by Polaris-B and Polaris-C to perform $m\log(m)$ iterations on different versions of the Walmart dataset (left), and the distribution of the relative errors of color assortativity of the samples generated by CM (right). Starting from the 11 available colors (product categories), we cluster these colors to create new realistic sets of 2, 4, and 8 colors.

The running time of CM is not affected by the number of colors, as it samples from a state space that is agnostic to vertex attributes. Therefore we omit it from this plot. Interestingly, the running time of Polaris-C remains consistent across the different numbers of colors. This consistency is likely due to Polaris-C maintaining a

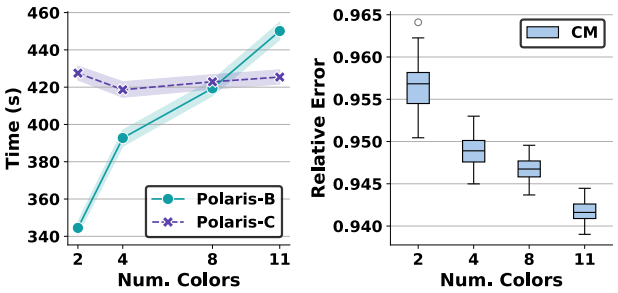


Figure 4: Running time (left) required by Polaris-B and Polaris-C to perform $m\log(m)$ iterations in different versions of Walmart, and distribution of the relative errors of color assortativity of the samples generated by CM (right). Results for 100 samples.

high acceptance rate, which produces a similar number of updates regardless of the number of colors.

In contrast, the running time of Polaris-B grows as the number of colors increases. As already mentioned, a higher number of colors increases the probability that a sampled DES is not a JDES. Consequently, more DESs need to be drawn before finding one that is a JDES, which leads to an increased running time.

The figure also shows the distribution of the relative error of color assortativity values for the multigraphs generated by CM. Again, we observe that the color assortativity of the sampled multigraphs significantly diverges from those of the original multigraphs.

5.2 Performance Analysis of Polaris-∗

Figure 5 consists of three panels, each addressing a different aspect of the performance and behavior of Polaris-B and Polaris-C, in three different datasets: Cite, Brexit, and Twitter. The top panel shows the running time as a function of the number of iterations. Polaris-C is faster than Polaris-B in most cases.

The middle panel shows the fraction of iterations with four possible outcomes: (*i*) the sampled DES is not a JDES (*Out of Space*), (*ii*) the DES is a no-op JDES (*Unchanged*), (*iii*) an accepted transition to the next state (*Accepted*), and (*iv*) a rejected transition (*rejected*). Polaris-C avoids sampling DESs that are not JDES, thus having no *Out of Space* outcomes. Additionally, Polaris-C has a higher ratio of accepted transitions than Polaris-B, thus it explores the state space more extensively. Due to frequent *Out of Space* outcomes, Polaris-B has to frequently resample new DESs, leading to increased running times per iteration (as shown in the first panel). This effect becomes particularly evident as the number of colors increases.

The bottom panel illustrates the *degree* assortativity of the states visited in the Markov chains produced by each sampler, which is often used as a measure of convergence for the chain [45]. Both algorithms reach a plateau at nearly the same point, which suggests the states visited start to share similar characteristics. However, as shown in the first plot, Polaris-C achieves this plateau faster.

Figure 6 provides a detailed analysis of the time required to perform a step in the sampling process, categorized by the type of outcome. Specifically, the left chart shows the times for transitions that are accepted, whereas the middle chart illustrates the times for transitions that are rejected. An accepted transition corresponds

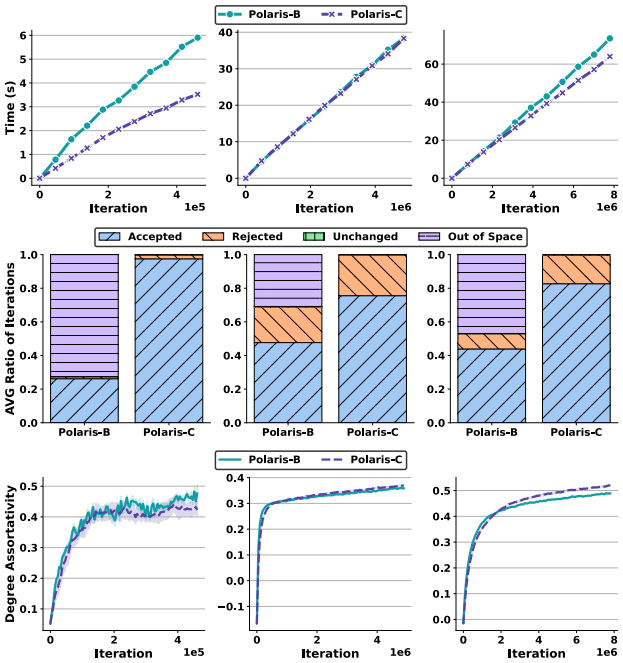
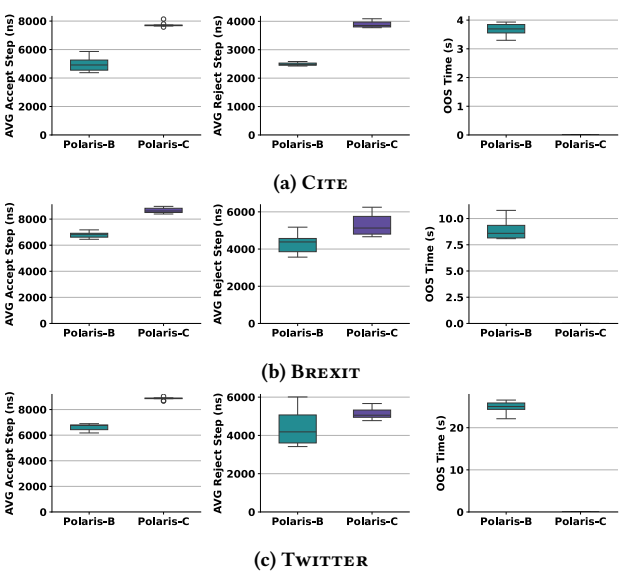


Figure 5: Running time (top), average ratio of iterations for each of the four possible outcomes (mid), and degree assortativity as a function of the number of iterations (bottom) for each sampler on Cite (left), Brexit (middle), and Twitter (right).



(a) Cite

(b) Brexit

(c) Twitter

Figure 6: Distribution of the average time required to perform a step where the transition to the next state is accepted (left) or rejected (middle). The right plots show the total time required to find a DES that is a JDES. Results for 10 Markov chains.

to the outcome *Accepted*, while a rejected transition include the outcomes *Rejected* and *Unchanged*. The right chart displays the distribution of total time for steps where the sampled DES is not a JDES (i.e., *Out of Space*). On average, accepting a transition is 2× slower than rejecting it, because the algorithms must update the data structures that store the edges and their weights to maintain correctness after an accepted transition.

Polaris-C has higher step times compared to Polaris-B as

- Polaris-C performs more computations even for rejected steps, as it must evaluate several quantities to compute the value of ρ. In contrast, Polaris-B performs fewer computations before rejection, and thus achieves lower running times;
- when a transition is accepted, Polaris-C needs to maintain additional data structures necessary to ensure that the sampled DES is always a JDES.

Nonetheless, Polaris-B results in longer overall running times because it uses a considerable amount of time to find a DES that is a JDES, especially when the number of colors is higher.

6 Conclusion

We introduced Polaris, an ensemble of colored multigraphs with prescribed Joint Color Matrix. The JCM captures key properties relevant to the study of polarized networks, such as color assortativity. We described two efficient algorithms to sample from the space of such multigraphs according to any user-specified probability distribution over this space. Our algorithms work by running a Markov chain on the multigraph space, following the Metropolis-Hastings approach for determining whether to accept the move to a proposed neighbor of the current state. We conducted an extensive experimental evaluation, showing the shortcomings of existing methods in capturing the color assortativity, and assessing the performance of our algorithms across different datasets in terms of scalability, runtime, and acceptance probability.

This work serves as an important first step toward analyzing polarization in real networks. While a comprehensive study of polarization lies beyond this paper's scope, the tools developed here lay the groundwork for future studies on this subject.

7 Ethical considerations

Our paper introduces a new method for assessing the statistical significance of the polarization structure discovered in online social networks. The motivation of our work is the study of social phenomena, such as polarization, formation of echo chambers, opinion dynamics, and influence among individuals in online social networks. As such, our work contributes to the growing field of computational social science, which in turn contributes to a better understanding of complex social behavior. Our emphasis in this work is on the design of new algorithms for efficient sampling from a novel network null model and the mathematical analysis of the algorithms and their properties. During our study, we did not perform any data collection, and our empirical evaluation uses benchmark graph datasets that are publicly available. For future studies and researchers who would like to apply our method on new data collected from online social networks, we emphasize the importance of prioritizing ethical considerations to protect the privacy and rights of individuals. This involves obtaining informed consent where possible, anonymizing data to prevent the identification of users, applying the data minimization principle, and being mindful of the potential for harm in the analysis and dissemination

of findings. It is also important to comply with platform policies and legal regulations, such as GDPR.

Acknowledgments

MR's work is supported by the National Science Foundation grants IIS-2006765 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=2006765), and CAREER-2238693 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=2238693). AG's work is supported by the ERC Advanced Grant REBOUND (834862), the EC H2020 RIA project SoBigData++ (871042), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Ilya Amburg, Nate Veldt, and Austin R. Benson. 2020. Clustering in graphs and hypergraphs with categorical edge labels. In *Proceedings of the Web Conference*. (Cited on 12)

[2] Yael Artzy-Randrup and Lewi Stone. 2005. Generating uniformly distributed random networks. *Physical Review E* 72, 5 (2005), 056708. (Cited on 3)

[3] Delia Baldassarri and Scott E. Page. 2021. The emergence and perils of polarization. *Proceedings of the National Academy of Sciences* 118, 50 (2021), e2116863118. https://doi.org/10.1073/pnas.2116863118 (Cited on 1)

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512. (Cited on 2)

[5] Edward A Bender and E.Rodney Canfield. 1978. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A* 24, 3 (May 1978), 296–307. https://doi.org/10.1016/0097-3165(78)90059-6 (Cited on 1, 2)

[6] Béla Bollobás. 1980. A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs. *European Journal of Combinatorics* 1, 4 (Dec. 1980), 311–316. https://doi.org/10.1016/S0195-6698(80)80030-8 (Cited on 1, 2)

[7] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. 2001. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms* 18, 3 (2001), 279–290. (Cited on 2)

[8] Sonia Cafieri, Pierre Hansen, and Leo Liberti. 2010. Loops and multiple edges in modularity maximization of networks. *Physical Review E* 81, 4 (2010), 046102. (Cited on 3)

[9] Philip S Chodrow, Nate Veldt, and Austin R Benson. 2021. Hypergraph clustering: from blockmodels to modularity. *Science Advances* (2021). (Cited on 12)

[10] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The Echo Chamber Effect on Social Media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118. https://doi.org/10.1073/pnas.2023301118 (Cited on 1)

[11] Edward F Connor and Daniel Simberloff. 1979. The assembly of species communities: chance or competition? *Ecology* 60, 6 (1979), 1132–1140. (Cited on 2)

[12] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*. 89–96. (Cited on 1, 12)

[13] Colin Cooper, Martin Dyer, and Catherine Greenhill. 2007. Sampling regular graphs and a peer-to-peer network. *Combinatorics, Probability and Computing* 16, 4 (2007), 557–593. (Cited on 2)

[14] Upasana Dutta, Bailey K. Fosdick, and Aaron Clauset. 2021. Sampling random graphs with specified degree sequences. https://doi.org/10.48550/ARXIV.2105.12120 Version Number: 4. (Cited on 2)

[15] Paul Erdős and Alfréd Rényi. 1959. On random graphs. *Publicationes Mathematicate* 6 (1959), 290–297. (Cited on 2)

[16] Péter L. Erdős, Catherine Greenhill, Tamás Róbert Mezei, István Miklós, Daniel Soltész, and Lajos Soukup. 2022. The mixing time of switch Markov chains: a unified approach. *European Journal of Combinatorics* 99 (2022), 103421. (Cited on 5)

[17] Ronald Aylmer Fisher. 1935. The design of experiments. Oliver and Boyd. *Edinburgh* (1935). (Cited on 2)

[18] World Economic Forum. 2024. *The Global Risks Report 2024*. Technical Report. https://www.weforum.org/publications/global-risks-report-2024 (Cited on 1)

[19] Bailey K. Fosdick, Daniel B. Larremore, Joel Nishimura, and Johan Ugander. 2017. Configuring Random Graph Models with Fixed Degree Sequences. arXiv:1608.00607 (Cited on 2, 4, 11)

[20] Bailey K. Fosdick, Daniel B. Larremore, Joel Nishimura, and Johan Ugander. 2018. Configuring Random Graph Models with Fixed Degree Sequences. *Siam Review* 60, 2 (2018), 315–355. (Cited on 1, 3, 4, 10)

[21] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 33–42. (Cited on 1)

[22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing Controversy by Connecting Opposing Views. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 81–90. (Cited on 1)

[23] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the World Wide Web Conference (WWW)*. 913–922. (Cited on 1, 12)

[24] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing* 1, 1 (2018), 3. (Cited on 1)

[25] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. *Advances in neural information processing systems* 30 (2017). (Cited on 12)

[26] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. The ebb and flow of controversial debates on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 524–527. (Cited on 12)

[27] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian data analysis*. Chapman and Hall/CRC. (Cited on 2)

[28] Matthew Gentzkow and Jesse M. Shapiro. 2011. Ideological Segregation Online and Offline. *The Quarterly Journal of Economics* 126, 4 (2011), 1799–1839. https://doi.org/10.1093/qje/qjr044 (Cited on 1)

[29] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet De Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* 381, 6656 (2023), 392–398. https://doi.org/10.1126/science.ade7138 (Cited on 1)

[30] Nicholas J. Gotelli and Gary R. Graves. 1996. *Null models in ecology*. Smithsonian Institution Press. (Cited on 3)

[31] Catherine Greenhill. 2014. The switch Markov chain for sampling irregular graphs. In *Proceedings of the 26th annual ACM-SIAM Symposium on Discrete Algorithms*. 1564–1572. (Cited on 2)

[32] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet De Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* 381, 6656 (2023), 404–408. https://doi.org/10.1126/science.add8424 (Cited on 1)

[33] Marilena Hohmann, Karel Devriendt, and Michele Coscia. 2023. Quantifying ideological polarization on a network using generalized Euclidean distance. *Science Advances* 9, 9 (2023), eabq2044. https://doi.org/10.1126/sciadv.abq2044 Publisher: American Association for the Advancement of Science. (Cited on 1)

[34] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444. (Cited on 1)

[35] Solomon Messing and Sean J. Westwood. 2014. Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research* 41, 8 (2014), 1042–1063. https://doi.org/10.1177/0093650212466406 (Cited on 1)

[36] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827. (Cited on 2)

[37] Michael Mitzenmacher and Eli Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press. (Cited on 3)

[38] Jacob L Moreno and Helen H Jennings. 1938. Statistics of social configurations. *Sociometry* (1938), 342–374. (Cited on 2)

[39] Mark Newman. 2018. *Networks*. Oxford university press. (Cited on 6)

[40] M. E. J. Newman. 2003. Mixing patterns in networks. *Physical Review E* 67, 2 (2003), 026126. https://doi.org/10.1103/PhysRevE.67.026126 (Cited on 1)

[41] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón,

Andrew M. Guess, Edward Kennedy, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Jennifer Pan, Daniel Robert Thomas, Rebekah Tromble, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong, Chad Kiewiet De Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620, 7972 (2023), 137–144. https://doi.org/10.1038/s41586-023-06297-w (Cited on 1)

[42] Peter H Peskun. 1973. Optimum monte-carlo sampling using markov chains. *Biometrika* 60, 3 (1973), 607–612. (Cited on 2, 5)

[43] Ivo Ponocny. 2001. Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika* 66 (2001), 437–459. (Cited on 11)

[44] Giulia Preti, Gianmarco De Francisci Morales, and Matteo Riondato. 2023. Maniacs: Approximate mining of frequent subgraph patterns through sampling. *ACM Transactions on Intelligent Systems and Technology* 14, 3 (2023), 1–29. (Cited on 12)

[45] Vivekananda Roy. 2019. Convergence diagnostics for Markov chain Monte Carlo. http://arxiv.org/abs/1909.11827 arXiv:1909.11827 [stat]. (Cited on 2, 7)

[46] Ali Salloum, Ted Hsuan Yun Chen, and Mikko Kivelä. 2022. Separating Polarization from Noise: Comparison and Normalization of Structural Polarization Measures. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–33. https://doi.org/10.1145/3512962 (Cited on 1)

[47] Tom Snijders, et al. 2002. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* 3, 2 (2002), 1–40. (Cited on 2)

[48] Tiziano Squartini and Diego Garlaschelli. 2017. *Maximum-entropy networks: Pattern detection, network reconstruction and graph combinatorics*. Springer. (Cited on 2)

[49] Lewi Stone and Alan Roberts. 1990. The checkerboard score and species distributions. *Oecologia* 85 (1990), 74–79. (Cited on 3)

[50] Richard Taylor. 2006. Contrained switchings in graphs. In *Proceedings of the Eighth Australian Conference on Combinatorial Mathematics*. Springer, 314–336. (Cited on 3)

[51] Petter Törnberg. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences* 119, 42 (2022), e2207159119. https://doi.org/10.1073/pnas.2207159119 (Cited on 1)

[52] Norman D. Verhelst. 2008. An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika* 73, 4 (2008), 705–728. (Cited on 3)

[53] Fabien Viger and Matthieu Latapy. 2005. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *International computing and combinatorics conference*. Springer, 440–449. (Cited on 13)

A Supplementary Material

A.1 Missing Proofs

In this section we present the proofs that were not included in the main body of the paper.

Proof of Thm. 4.3. For any two multigraphs $Q, R \in \mathcal{Z}$, and any vertex $u \in V$, define

$$\Delta_{Q,R}(u) \doteq \sum_{\ell \in \mathcal{L}} \left| \gamma_Q^\ell(u) - \gamma_R^\ell(u) \right|,$$

i.e., as the sum of the absolute differences in the numbers of neighbors of u with the same color across Q and R. Also define

$$\Delta(Q,R) \doteq \sum_{u \in V} \Delta_{Q,R}(u) \ .$$

Let G and H be two distinct multigraphs in $\mathcal{Z}$. We first build a sequence of JDESs to transform G into a multigraph $\tilde{G} \in \mathcal{Z}$ such that $\Delta(\tilde{G}, H) = 0$, i.e., every vertex has in $\tilde{G}$ exactly the same number of neighbors of each color as it has in H. Then, we construct a sequence of JDESs to transform $\tilde{G}$ into H.

If $\Delta(G,H) = 0$, then let $\tilde{G} = G$. Otherwise, let u be a vertex such that $\Delta_{G,H}(u) > 0$. From this fact and since $d_G(u) = d_H(u)$ (as the two multigraphs have the same degree sequence), then, there are distinct $j, k \in \mathcal{L}$ s.t. $\gamma_G^j(u) < \gamma_H^j(u)$ and $\gamma_G^k(u) > \gamma_H^k(u)$.

It follows from the first inequality and the fact that G and H have the same JCM, that there must be a vertex $v \neq u$ with $\lambda(v) = \lambda(u)$ and such that $\gamma_G^j(v) > \gamma_H^j(v)$.

Let $w \in \Gamma_G^k(u)$ and $z \in \Gamma_G^j(v)$. These two vertices necessarily both exist because $\gamma_G^k(u) > \gamma_H^k(u) \geq 0$, and $\gamma_G^j(v) > \gamma_H^j(v) \geq 0$.

Since $\lambda(u) = \lambda(v)$, we can construct the JDES $(u,w),(v,z) \to (u,z),(v,w)$ that transforms G into some $T \in \mathcal{Z}$. This JDES falls in case 4C of the characterization of JDESs. It holds:

- $\Delta_{T,H}(u) = \Delta_{G,H}(u) - 2$;
- $\Delta_{T,H}(v)$ is either equal to $\Delta_{G,H}(v)$ or to $\Delta_{G,H}(v) - 2$, as $\left|\gamma_T^\ell(v) - \gamma_H^j(v)\right| = \left|\gamma_G^j(v) - \gamma_H^j(v)\right| - 1$, and $\left|\gamma_T^k(v) - \gamma_H^k(v)\right|$ is $\left|\gamma_G^k(v) - \gamma_H^k(v)\right| \pm 1$;
- $\Delta_{T,H}(q) = \Delta_{G,H}(q)$ for every other vertex $q \in V \setminus \{u,v\}$, as w and z exchanged neighbors with the same color $\lambda(u)$, and every other vertex has no change in its neighborhood.

Therefore, it holds $\Delta(T,H) \leq \Delta(G,H) - 2$. By repeatedly applying the procedure above, we eventually obtain a graph $\tilde{G} \in \mathcal{Z}$ such that $\Delta(\tilde{G}, H) = 0$.

Now, if $\tilde{G} = H$, we are done. Otherwise, for any multigraph $Q = (V, E, \mathcal{L}, \lambda) \in \mathcal{Z}$, and any unordered pair (ℓ, r) of colors from $\mathcal{L}$ (potentially $\ell = r$), define $V_{\ell,r} = \{v \in V : \lambda(v) \in \{\ell, r\}\}$, and define the subgraph $Q_{\ell,r}$ of Q as the multigraph

$$Q_{\ell,r} = (V_{\ell,r}, \{\!\{(u,v) \in E : \lambda(u) = \ell \wedge \lambda(v) = r\}\!\}, \{\ell, r\}, \lambda_{|V_{\ell,r}}),$$

i.e., the subgraph of Q that contains only the vertices with color ℓ or r, and only the edges that have one endpoint with color ℓ and one endpoint with color r. Clearly, if $\ell = r$, $Q_{\ell,\ell}$ is the subgraph of Q induced by the vertices with color ℓ, but if $\ell \neq r$, $Q_{\ell,r}$ is not the subgraph of Q induced by the vertices with color ℓ or r, as in this case $Q_{\ell,r}$ does not include any eventual self-loop over its vertices, as its edges are all and only the bichrome edges with one vertex of color ℓ and the other vertex of color r. The edge sets of the various $Q_{\ell,r}$ form a partitioning of the edge set E of Q.

For $\ell \in \mathcal{L}$, consider $\tilde{G}_{\ell,\ell}$ and $H_{\ell,\ell}$. These multigraphs have the same set of vertices, and the same degree sequence, as $\Delta(\tilde{G}, H) = 0$. Since their vertices all have color ℓ, any DES from $\tilde{G}_\ell$ is a JDES, falling in either case 1, 2A, 2B, 3A, 3B, or 4C. A classic result of graph theory (see, e.g., [20, Lemma 2.14]) states that there is a sequence of double edge swaps connecting any multigraph with vertices all with the same color to any other multigraph with the same degree sequence as the starting multigraph. Thus there is a sequence of JDESs that connects $\tilde{G}_{\ell,\ell}$ to $H_{\ell,\ell}$. For any arbitrary ordering $\ell_1, \ell_2, \ldots, \ell_{|\mathcal{L}|}$ of the colors in $\mathcal{L}$, we can then consider the sequence of JDESs obtained by concatenating the sequences of JDESs connecting $\tilde{G}_{\ell_i,\ell_i}$ to H_{ℓ_i,ℓ_i}, and apply the JDESs in the resulting sequence, starting from $\tilde{G}$, to obtain $\check{G}$ with $\check{G}_{\ell,\ell} = H_{\ell,\ell}$ for $\ell \in \mathcal{L}$.

The same approach can also be used for $\check{G}_{\ell,\mathrm{r}}$ $(= \tilde{G}_{\ell,r})$ and $H_{\ell,r}$ for $\ell \neq r$. They have the same set of vertices and the same degree sequences, and any DES is a JDES, falling in either case 2C, 3D, or 4A. Every bipartite multigraph can be transformed, through a sequence of DES, into any other bipartite multigraph with the same set of vertices and the same degree sequence. Thus there is a sequence of JDESs that transforms $\tilde{G}_{\ell,\mathrm{r}}$ $(= \tilde{G}_{\ell,r})$ into $H_{\ell,r}$, for $\ell \neq r$. Using the same arbitrary ordering $\ell_1, \ell_2, \ldots, \ell_{|\mathcal{L}|}$ of the colors in $\mathcal{L}$, we can then consider the ordering $(\ell_1, \ell_2), (\ell_1, \ell_3), \ldots, (\ell_1, \ell_{|\mathcal{L}|}), (\ell_2, \ell_3), \ldots, (\ell_{|\mathcal{L}|-1}, \ell_{|\mathcal{L}|})$ of the unordered pairs of different colors,

and consider the sequence of JDESs obtained by concatenating the sequences of JDESs connecting $\check{G}_{\ell_i,\ell_j}$, $i < j$, to H_{ℓ_i,ℓ_j}, to obtain H.

We have thus built a sequence of JDESs that starts at G, goes through $\tilde{G}$ and $\check{G}$, and reaches H. Since every JDES is reversible, our proof is complete. □

Proof of Thm. 4.4. Assume that only the first condition holds.

If the edges fall in Cases 1, 2B, 2C, 2D, 3B, 3C, 3D, or 3E, then at least one of the JDESs involving them is a no-op, so the state graph has a self-loop on state G, and is therefore aperiodic.

If the edges fall in Cases 2A or 3A, then applying either of the equivalent JDESs involving the edges leads to a state $H \neq G$ where the new edges fall in case 2B or either 3B or 3C respectively, meaning that the state graph has a self-loop on H, thus it is aperiodic.

If the edges fall in case 4C,[4] the state graph has a cycle of length two, because every JDES is reversible. It also has a cycle of length three, by applying the following sequence of JDESs: $(u,w),(v,z) \to (u,v),(w,z)$, $(u,v),(w,z) \to (u,z),(v,w)$, and $(u,z),(v,w) \to (u,w),(v,z)$. The greatest common divisor of the lengths of these two cycles is one, thus the state graph is aperiodic.

Assume now that only the second condition holds.[5] The state graph has a cycle of length two, because every JDES is reversible. It also has a cycle of length three, by applying the following sequence of JDESs: $(u,z),(w,x) \to (u,x),(w,z)$, $(u,v),(w,z) \to (u,z),(w,v)$, $(u,x),(w,v) \to (u,v),(w,x)$. The greatest common divisor of the lengths of these two cycles is one, thus the state graph is aperiodic. □

Proof of Thm. 4.5. As shown in Thms. 4.3 and 4.4, the state graph is strongly-connected and aperiodic. For every $G, H \in \mathcal{Z}$ s.t. H is a neighbor of G, Polaris-B the probability $\xi_G(H)$ of proposing H when the state of the chain is G is the same as in [19, Algorithm 3], and thus so is the ratio $\rho \doteq \xi_H(G)/\xi_G(H)$ used by Polaris-B, and therefore the acceptance probability $\alpha_G(H) \doteq \min\{1, \rho\pi(H)/\pi(G)\}$. Thus, Polaris-B follows the MH approach, and the Markov chain it runs has stationary distribution π. □

We now give the proof to Thm. 4.6. In the proof, we use the following immediate facts.

Fact A.1. *For any $\ell \in \mathcal{L}$ and any pair of distinct edges (u,w), $(v,z) \in E_{G,\ell}$, there is always a JDES from G involving these two edges, as $\{\lambda(u),\lambda(w)\} \cap \{\lambda(v),\lambda(z)\} \neq \emptyset$, i.e., Case 0 in the characterization of JDESs never holds.*

Fact A.2. *Let $(u,w),(v,z) \in E_{G,\ell}$. If $|\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\}| \neq 2$, there exists no $r \in \mathcal{L}$, $r \neq \ell$, such that $(u,w),(v,z) \in E_{G,r}$. Otherwise, there is exactly one such r, as $\{\lambda(u),\lambda(w),\lambda(v),\lambda(z)\} = \{\ell, r\}$.*

Fact A.3. *If $G, H \in \mathcal{Z}$ are neighbors and the one or two JDESs that transform G into H fall in case XY from the classification above, then the one or two JDESs that transform H into G fall in the same case XY, with the following exceptions:*

XY=2A: *the only JDES from H to G falls in case 2B;*
XY=2B: *the two equivalent JDESs from H to G fall in case 2A;*
XY=3A: *the only JDES from H to G falls in either case 3B or 3C, depending on whether the three involved vertices have all the same color (3B) or not (3C);*
XY=3B: *the only JDES from H to G falls in case 3A;*
XY=3C: *the only JDES from H to G falls in case 3A.*

Fact A.4. *Let $G \neq H \in \mathcal{Z}$. For any $\ell \in \mathcal{L}$, it holds $|E_{G,\ell}| = |E_{H,\ell}|$, i.e., the size of these sets is constant across all multigraphs in $\mathcal{Z}$.*

The above fact does not imply $E_{G,\ell} = E_{H,\ell}$.

Proof of Thm. 4.6. Let $G \in \mathcal{Z}$ be the current state of the Markov Chain, i.e., the value taken by the variable G at the beginning of some iteration of the loop (line 2). We aim to show that Polaris-C follows the Metropolis-Hastings (MH) approach. To this end, we need to show that the acceptance probability $\alpha_G(H)$ is computed according to the MH approach, for any neighbor H of G.

Let ℓ take value $r \in \mathcal{L}$ (line 3), and let $(u,w) = (a,b)$ and $(v,z) = (c,d)$ be the edges sampled from $E_{G,r}$ (lines 4–5). It holds from Fact A.1 that here is always a JDES from G involving these edges (i.e., Case 0 from the characterization never holds). Let $(a,b),(c,d) \to (a,c),(b,d)$ be a JDES involving these edges, and let $H \in \mathcal{Z}$ be the multigraph resulting from applying the JDES to H. We now consider the different cases for the JDES.

In Cases 1, 2C, 2D, 3D, or 3E: it holds $H = G$, and indeed the algorithm does not update G (lines 6, 15, and 28), i.e., the state of the Markov chain is unchanged, as required by MH.

For all other cases, it will be $H \neq G$. To assess $\alpha_G(H)$, we need to study the neighbor proposal probabilities $\xi_G(H)$ and $\xi_H(G)$. More specifically, we study the value of the variable ρ set by Polaris-C, and we show that is always set to $\xi_H(G)/\xi_G(H)$. The correctness of the algorithm then follows from this fact and the fact that Polaris-C decides whether to accept H by comparing a real drawn uniformly at random from $[0,1]$ to the value $\rho\pi H/\pi G$, i.e., it accepts H with probability $\alpha_G(H) = \min\{1, \rho\pi H/\pi G\}$, as required by MH.

Consider the random variables ℓ, (u,w), (v,z), and jdes used by the algorithm, and defined the following events:

$\mathsf{E_l}$: $\ell = r$ such that $(a,b),(c,d) \in E_{G,r}$;
$\mathsf{E_e}$: $((u,w) = (a,b) \wedge (v,z) = (c,d)) \vee ((u,w) = (c,d) \wedge (v,z) = (a,b))$;
$\mathsf{E_j}$: jdes $= (a,b),(c,d) \to (a,c),(b,d)$.

It clearly holds $\xi_G(H) = \Pr(\mathsf{E_j})$. Using the law of total probability, we can write:

$$\xi_G(H) = \Pr(\mathsf{E_j}) = \Pr(\mathsf{E_l}) \Pr(\mathsf{E_e} \mid \mathsf{E_l}) \Pr(\mathsf{E_j} \mid \mathsf{E_e}) \ .$$

We now analyze jdes, using the characterization of the JDESs.

Case 3A, 3B, 3C, 4A: for these cases, it holds $\{\lambda(a),\lambda(b)\} \cap \{\lambda(c),\lambda(d)\} = \{r\}$ for some $r \in \mathcal{L}$. Thus, Fact A.2 tells us that (a,b) and (c,d) appear together only in $E_{G,r}$, hence $\Pr(\mathsf{E_l}) = 1/|\mathcal{L}|$. Also, the two edges are not copies of the same multiedge, hence

$$\Pr(\mathsf{E_e} \mid \mathsf{E_l}) = \frac{\omega_G((a,b))\omega_G((c,d))}{|E_{G,\ell}|(|E_{G,\ell}| - 1)} \ . \tag{1}$$

Finally, it clearly holds $\Pr(\mathsf{E_j} \mid \mathsf{E_e}) = 1$. Thus,

$$\xi_G(H) = \frac{\omega_G((a,b))\omega_G((c,d))}{|\mathcal{L}||E_{G,\ell}|(|E_{G,\ell}| - 1)} \ . \tag{2}$$

[4] The proof for this case is the same as the proof for the aperiodicity of the state graph by DESs [19, Lemma 3].
[5] The proof for this case is the same as the proof for the aperiodicity of the state graph by DES when considering only bipartite graphs as the states [43, Sect. 4.2.1].

We know from Fact A.3 that the JDES $(a,c),(b,d) \to (a,b),(c,d)$ from H to G would fall in a case among those we are considering here, so we can obtain $\xi_H(G)$ from eq. (2) as

$$\xi_H(G) = \frac{\omega_H((a,c))\omega_H((b,d))}{|\mathcal{L}||E_{H,\ell}|(|E_{H,\ell}|-1)} \ .$$

We can then use Fact A.4 and the fact that $\omega_H((a,c)) = \omega_G((a,c))+1$ and $\omega_H((b,d)) = \omega_G((b,d))+1$ to rewrite the above as

$$\xi_H(G) = \frac{(\omega_G((a,c))+1)(\omega_G((b,d))+1)}{|\mathcal{L}||E_{G,\ell}|(|E_{G,\ell}|-1)} \ . \tag{3}$$

It follows that Polaris-C, on lines 20, 24, 27 and 33 sets ρ to the ratio $\xi_H(G)/\xi_G(H)$, as requested.

Case 4C: in this case, at least one of the edges is monochrome, so $\Pr(\mathsf{E}_\mathsf{l}) = 1/|\mathcal{L}|$. The edges are not copies, so $\Pr(\mathsf{E}_\mathsf{e})$ is as in eq. (1). But it holds $\Pr(\mathsf{E}_\mathsf{j} \mid \mathsf{E}_\mathsf{e}) = 1/2$, since jdes is equally likely to be either of the two JDESs involving the sampled edges, only one of which transforms G into H. Thus, $\xi_G(H)$ equals the r.h.s. of eq. (2) multiplied by $1/2$. The JDES from H to G falls in this same case, so we can proceed as in the previous case, and $\xi_H(G)$ equals the r.h.s. of eq. (3) multiplied by $1/2$. Thus, Polaris-C sets ρ on lines 41 and 44 to the ratio $\xi_H(G)/\xi_G(H)$, as required.

Case 2A: this case is partly similar to Cases 3A, 3B, 3C, 4A, and the value for $\xi_G(H)$ is the same as in eq. (2), noting that, in this case, the JDES from G to H can be written as $(a,a),(c,c) \to (a,c),(a,c)$. On the other hand, we know from Fact A.3 that the JDES $(a,c),(a,c) \to (a,a),(c,c)$ from H to G falls in case 2B. This case is analyzed next, and the proposal probability $\xi_H(G)$ can be obtained from eq. (4). By using Fact A.4 and the fact that $\omega_H((a,c)) = \omega_G((a,c))+2$, similarly to how we proceeded in the above cases, we obtain

$$\xi_H(G) = \frac{(\omega_G((a,c))+2)(\omega_G((a,c))+1)}{|\mathcal{L}||E_{G,\ell}|(|E_{G,\ell}|-1)} \ .$$

Thus, Polaris-C sets ρ on line 10 to the ratio $\xi_H(G)/\xi_G(H)$, as required.

Case 2B: this case is also partly similar to the first we analyzed, except that the two edges are copies of the same multiedge (a,b), hence

$$\Pr(\mathsf{E}_\mathsf{e} \mid \mathsf{E}_\mathsf{l}) = \frac{\omega_G((a,b))(\omega_G((a,b))-1)}{|E_{G,\ell}|(|E_{G,\ell}|-1)}$$

and

$$\xi_G(H) = \frac{\omega_G((a,b))(\omega_G((a,b))-1)}{|\mathcal{L}||E_{G,\ell}|(|E_{G,\ell}|-1)} \ . \tag{4}$$

From Fact A.3, we know that the JDES from H to G would fall in case 2A, thus the proposal probability $\xi_H(G)$ can be obtained from eq. (2), following a process similar to the one we described in case 2A, and resulting in

$$\xi_H(G) = \frac{(\omega_G((a,a))+1)(\omega_G((b,b))+1)}{|\mathcal{L}||E_{G,\ell}|(|E_{G,\ell}|-1)} \ .$$

Therefore, Polaris-C sets ρ on line 14 to the ratio $\xi_H(G)/\xi_G(H)$, as required.

Case 4B: in this case, the two edges (a,b) and (c,d) are both bichrome with the same two colors r' and r'', thus Fact A.2 tells us that they appear together in both $E_{G,r'}$ and $E_{G,r''}$, hence $\Pr(\mathsf{E}_\mathsf{l}) = 2/|\mathcal{L}|$. The edges are not copies, so

$$\Pr(\mathsf{E}_\mathsf{e} \mid \mathsf{E}_\mathsf{l}) = \frac{\omega_G((a,b))\omega_G((c,d))}{|E_{G,r'}|(|E_{G,r'}|-1)} + \frac{\omega_G((a,b))\omega_G((c,d))}{|E_{G,r''}|(|E_{G,r''}|-1)} \ .$$

It holds $\Pr(\mathsf{E}_\mathsf{j} \mid \mathsf{E}_\mathsf{e}) = 1$, thus

$$\xi_G(H) = \frac{2}{|\mathcal{L}|}\left(\frac{\omega_G((a,b))\omega_G((c,d))}{|E_{G,r'}|(|E_{G,r'}|-1)} + \frac{\omega_G((a,b))\omega_G((c,d))}{|E_{G,r''}|(|E_{G,r''}|-1)}\right) \ .$$

The JDES from H to G also falls in case 4B, per Fact A.3. Using Fact A.4, and the multiplicities of the edges in G to express those in H, we obtain

$$\xi_H(G) = \frac{2}{|\mathcal{L}|}\left(\frac{(\omega_G((a,d))+1)(\omega_G((c,b))+1)}{|E_{G,\ell}|(|E_{G,\ell}|-1)} + \frac{(\omega_G((a,d))+1)(\omega_G((c,b))+1)}{|E_{G,\ell'}|(|E_{G,\ell'}|-1)}\right) \ .$$

Once again, Polaris-C clearly sets ρ to the ratio $\xi_H(G)/\xi_G(H)$ on line 37. □

A.2 Datasets

We consider 11 real-world labeled networks, whose characteristics are summarized in Table 1. Brexit, US-Elect, Abortion [25], Twitter [12], Obamacare [26], Comb, and Guns [23] are retweet networks generated from tweets collected on various controversial topics. An edge exists between two users if one retweeted the other. Node colors indicate the side taken in the discussion, with a third label indicating neutrality. Cite and Phy-Cit [44] are citation networks: nodes represent publications, with node colors indicating Computer Science areas and the year of publication, respectively. Trivago [9] is network where nodes are accommodations, and edges connect accommodations visited by a user in the same browsing session. Node colors indicate the country where the accommodation is located. Walmart [1] is a co-purchase network where nodes are Walmart products, and edges connect products that were bought together. Node colors indicate the departments in which the products appear on walmart.com.

Table 1: Dataset characteristics: number of vertices, number of edges, average and median degree, and average and median color frequency.

Dataset	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert\mathcal{L}\rvert$	$\widehat{d(u)}$	$\widetilde{d(u)}$	$\widehat{\lvert V^\ell\rvert}$	$\widetilde{\lvert V^\ell\rvert}$
Cite	3264	4611	6	2.83	2.00	0.17	0.18
Brexit	22745	48830	2	4.29	1.00	0.50	0.50
Twitter	22405	77920	3	6.96	1.00	0.33	0.32
Phy-Cit	30501	347268	11	22.77	14.00	0.09	0.11
Abortion	279505	671144	2	4.80	1.00	0.50	0.50
US-Elect	23832	845152	3	70.93	3.00	0.33	0.25
Trivago	172738	1327092	160	15.37	6.00	0.01	0.00
Obamacare	334617	1511670	2	9.04	1.00	0.50	0.50
Walmart	88860	2267396	11	51.03	18.00	0.09	0.05
Comb	677753	6666018	2	19.67	1.00	0.50	0.50
Guns	632659	7478993	2	23.64	1.00	0.50	0.50

A.3 Computational Complexity

Polaris-C has an initialization step for the data structures used to ensure we only sample pairs of edges that form a JDES. This initialization phase has a time complexity of $O(n+m)$ and requires $O(n+m)$ space, where n is the number of vertices and m is the number of edges. The space is used to store information such as node degrees, node colors, the edge list, and the edge subsets $E_{G,\ell}$ for each color ℓ. Since each edge can belong to up to two of these subsets, they occupy at most $2m$ space in total. Each step of the algorithm then takes constant time, and we perform s steps. In our experiments, we follow previous works and set $s = m\log(m)$ as [53]. The number of colors $|\mathcal{L}|$ does not affect the time complexity, as Polaris-C directly samples pairs of edges that can form a JDES. In contrast, Polaris-B is affected by $|\mathcal{L}|$. As $|\mathcal{L}|$ increases, so does the number of possible combinations of node colors, thus reducing the probability that two sampled edges belong to the same subset $E_{G,\ell}$. Each time this condition is not met, Polaris-B must resample two new edges, which increases its running time.

geographical analysis

Check for updates

Geographical Analysis (2023) **55**, 239–267

Automated Detection of Missing Links in Bicycle Networks

Anastassia Vybornova[1], Tiago Cunha[1], Astrid Gühnemann[2], and Michael Szell[1,3,4]

[1]NEtwoRks, Data, and Society (NERDS), Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark, [2]Institute for Transport Studies, University of Natural Resources and Life Sciences, Vienna, Austria, [3]ISI Foundation, Turin, Italy, [4]Complexity Science Hub Vienna, Vienna, Austria

Cycling is an effective solution for making urban transport more sustainable. However, bicycle networks are typically developed in a slow, piecewise process that leaves open a large number of gaps, even in well-developed cycling cities like Copenhagen. Here, we develop the IPDC procedure (Identify, Prioritize, Decluster, Classify) for finding the most important missing links in urban bicycle networks, using data from OpenStreetMap. In this procedure we first identify all possible gaps following a multiplex network approach, prioritize them according to a flow-based metric, decluster emerging gap clusters, and manually classify the types of gaps. We apply the IPDC procedure to Copenhagen and report the 105 top priority gaps. For evaluation, we compare these gaps with the city's most recent Cycle Path Prioritization Plan and find considerable overlaps. Our results show how network analysis with minimal data requirements can serve as a cost-efficient support tool for bicycle network planning. By taking into account the whole city network for consolidating urban bicycle infrastructure, our data-driven framework can complement localized, manual planning processes for more effective, city-wide decision-making.

Introduction

With transport being one of the most problematic sectors in terms of emission reductions (Lamb et al. 2021), urban transportation systems play a decisive role in tackling the climate crisis. There is enormous potential to be harnessed by "greening" the transportation sector through a modal shift towards active and more sustainable mobility modes such as cycling and walking, both in terms of climate change mitigation and socioeconomic benefits (High-level Advisory Group on Sustainable Transport 2016; Gössling et al. 2019).

In practice, however, bicycle infrastructure development struggles with a particularly pervasive political inertia due to the complex interdependencies of car-centrism (Feddes, de Lange, and te Brömmelstroet 2020; Mattioli et al. 2020). Very few cities have so far managed to build up relatively safe and cohesive bicycle networks (de Groot 2016), and even the most renowned

Correspondence: Michael Szell, Computer Science Department, IT University of Copenhagen, 2300 Copenhagen, Denmark
e-mail: misz@itu.dk

Submitted: 10 January 2022; Revised version accepted: 2 March 2022

doi: 10.1111/gean.12324

cycling cities in the world still have a long way to go to achieve a sustainable urban transport system, and an optimal cycling network. For instance, this is the case for Copenhagen, where despite over a century of political struggles and coordinated efforts to develop a functioning grid of protected on-street bicycle networks (Carstensen et al. 2015), its network of protected bicycle infrastructure is split into 300 disconnected components (Natera Orozco et al. 2020a) and its accessibility displays considerable local variations (Rahbek 2020). For the assessment of an urban bicycle network, it is therefore crucial to ask: "Where are the missing links?", "How to fix them?", and "How much will this cost and benefit the city?" These are the questions we aim to answer in this paper. Our approach is based on Vybornova (2021) to develop a generally applicable, computational procedure for finding missing links in developed bicycle networks, and testing it on the case of Copenhagen.

From a research perspective, a structured, data-driven approach to bicycle network planning, along with a strong theoretical and computational underpinning, is largely missing. Setting up such an approach is seen by many as necessary precondition for an evidence-based modal shift towards increased bicycle use and reduced car use (Koglin and Rye 2014; Buehler and Dill 2016; de Groot 2016; Priya Uteng and Turner 2019; Resch and Szell 2019). From this viewpoint, the academic literature on network analysis approaches to bicycle network planning can be divided into three broad categories, depending on the structure and reproducibility of the underlying approach.

The first, largest category contains transport planning studies with a *place-specific* focus. These case studies focus on improving the bicycle infrastructure of one particular city, for example, Seattle (Lowry and Loh 2017), Toronto (Mitra, Ziemba, and Hess 2017), Lisbon (Abad and Van der Meer 2018) or London (Palominos, Smith, and Griffiths 2021). Characteristic for these studies is the specific application to one city and its idiosyncrasies, using a variety of data sets, such as orography, traffic flows, trip tables or citizen surveys on mobility preferences. The second, more recent approach, is based on the physics-inspired Science of Cities (Batty 2013) and aims to identify the generalized laws and mechanisms that govern urban development and are *independent of place*. This approach typically focuses on the most important "first-order" effects following the paradigm of network science, sacrificing specificity for generality, and therefore deliberately using maximally simplified data sets. Given that this second approach aims for general results, it must be tested for multiple cities. Examples include a multiplex network study of multimodality (Natera Orozco et al. 2020b), methods to prioritize pop-up active transport infrastructure (Lovelace et al. 2020), linking disconnected components (Natera Orozco et al. 2020a), or growing bicycle networks from scratch (Szell et al. 2021). Finally, the third category contains studies that develop *generalizable* approaches based on the use case of one specific city (Larsen, Patterson, and El-Geneidy 2013; Zhang, Magalhães, and Wang 2014; Boisjoly, Lachapelle, and El-Geneidy 2020; Olmos et al. 2020; Reggiani et al. 2021). There is an inherent feasibility trade-off between developing a refined model by working with one high resolution data set versus developing a generalizable model by working with several lower resolution data sets. Studies from the third category therefore often imply a call to the cycling research community to collaborate on method consolidation by further testing their respective approach for other cities.

Our approach developed here corresponds to the third category: we first develop a generalizable method for the detection of gaps in bicycle networks, and then carry out a detailed evaluation procedure for the use case of Copenhagen in order to demonstrate the applicability of our method. Our procedure should be applicable to other cities without major adjustments. Furthermore, this new procedure could also be applied to less developed networks, for example

to complement previous approaches (Fig. 1) or to find missing links in sub-networks below the scale of the city.

Lastly, there are also numerous approaches to bicycle network planning that focus on (actual or estimated) travel demand, often rooted in transport modeling (Dill and Gliebe 2008; Lovelace et al. 2017; Cooper 2018; Skov-Petersen et al. 2018; van Eldijk et al. 2020). The approaches divided in three categories above, in contrast, are all rooted in network analysis and focus on improvements of existing infrastructure, which our study is also in line with.

Our method complements Natera Orozco et al. (2020a) and Szell et al. (2021), see Fig. 1: Instead of providing optimized improvements to cities with minimal existing networks (Szell et al. 2021), or to cities with developed but still quite disconnected networks (Natera Orozco et al. 2020a), here we focus on repairing networks. This new approach is particularly suited for well-developed networks in which the largest connected components already cover the majority of nodes. These networks do not benefit from an approach that starts from scratch. They can benefit from connecting existing components (Natera Orozco et al. 2020a), but since they cover already most of the city, this benefit becomes exhausted quickly once the few biggest components have been linked up. However, there can still be many missing links left *within* their connected components, for which we set up an automated fixing procedure here, see Fig. 2.

The IPDC procedure

A cyclist on their way through an urban bicycle network will often find themselves suddenly having to share the road with cars for a while, or having to cross unprotected intersections with a

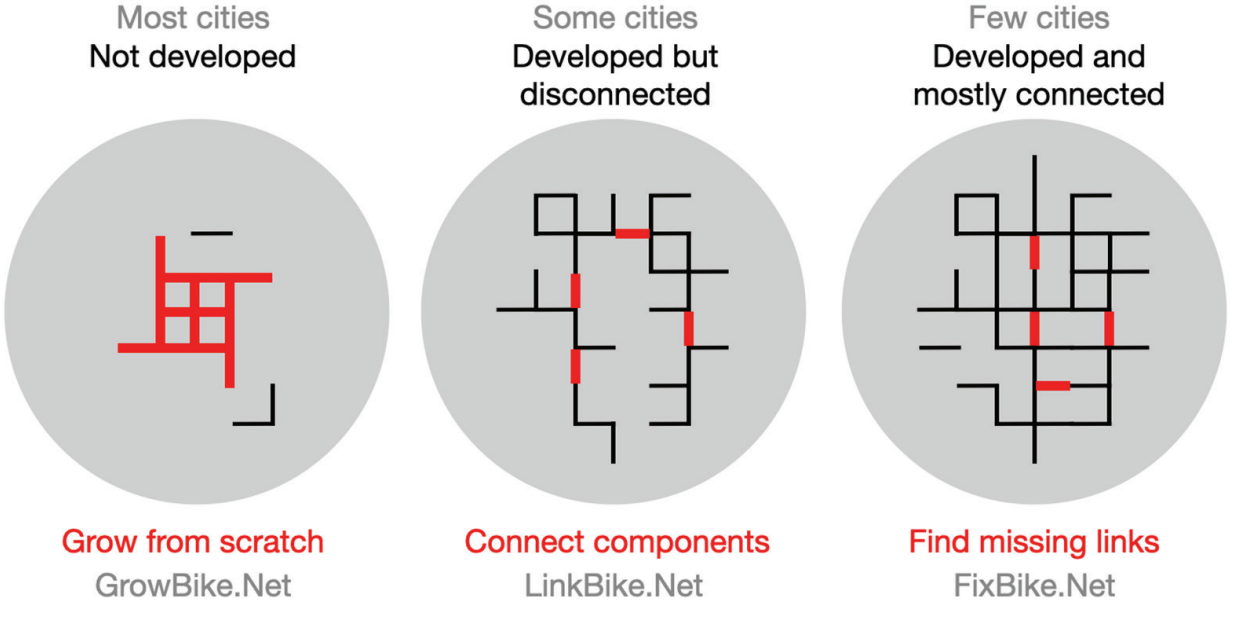


Figure 1. Depending on a city's existing bicycle network (black), different development approaches (red) can fit. Left: The approach of growing from scratch by Szell et al. (2021) is best applicable for underdeveloped cities, such as Los Angeles. See also: https://growbike.net. Center: The approach of Natera Orozco et al. (2020a) to connect disconnected components can well fit cities that have developed but disconnected components, such as Budapest. See also: http://linkbike.net. Right: Here we develop a process for finding missing links also within a connected component. This method complements the other two approaches; also it fits well cities with a developed and connected network such as Copenhagen. See also: https://fixbike.net. [Colour figure can be viewed at wileyonlinelibrary.com]

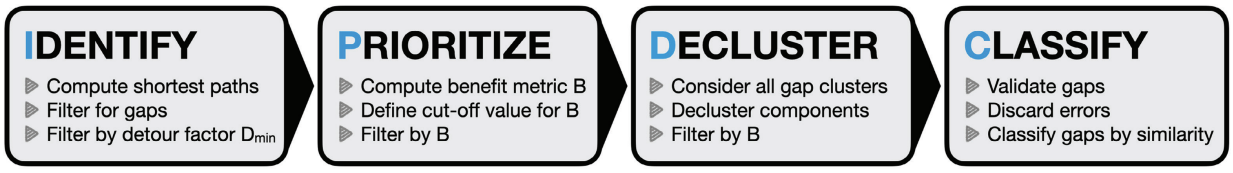


Figure 2. Steps in the IPDC procedure. Gaps are first identified via shortest paths, discarding parallel paths using a minimum detour factor $D_{\min}$. Gaps are then prioritized via a gap closure benefit metric, in the simplest case based on betweenness centrality. Resulting gaps can overlap (cluster) and need to be declustered. Finally, gaps are compared with existing infrastucture, validated or discarded, and classified.

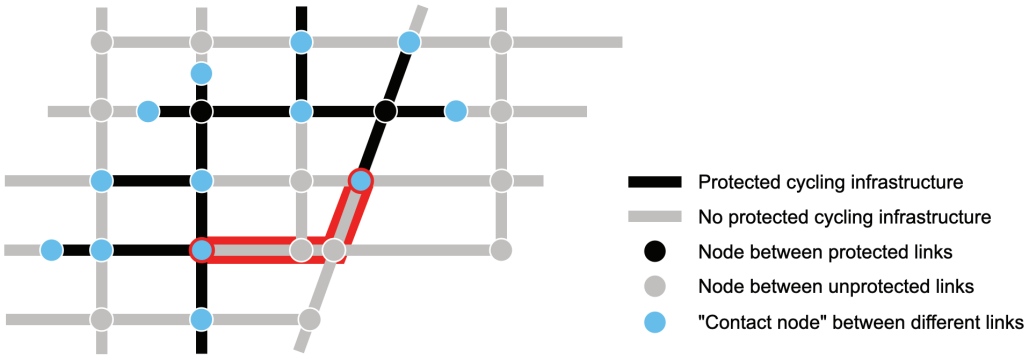


Figure 3. Illustration of node and link types, and of our definition of gap. We define a gap as a shortest path between two contact nodes that consists only of unprotected links. An example of a gap between the two highlighted contact nodes is illustrated in red. (Here we only illustrate one out of many possible gaps between all pairs of contact nodes.) [Colour figure can be viewed at wileyonlinelibrary.com]

high traffic load, even in a well-connected, developed bicycle network like in Copenhagen. Here we formalize this intuitive concept of a "missing link" in the bicycle network and develop an automated procedure to find the most important ones. We call our procedure *IPDC* after its four main steps: Identify, Prioritize, Decluster, Classify, which we present in this section. The IPDC procedure is illustrated in the workflow diagram in Fig. 2, and described in detail in the sections below. See Section *Scope and limitations* for details on the applicability and limitations of this approach. We start by outlining the network data structure and our formal definition of "gap" used for the first step of the IPDC procedure, gap identification.

Gap identification

As a starting point, the IPDC procedure takes an urban network of streets and protected bicycle tracks, as provided by OpenStreetMap (OSM). The steps to obtain and process the data are described in detail in Appendix A. The data are structured as a multiplex network (Battiston, Nicosia, and Latora 2014) with two different link types and three different node types, see Fig. 3. Links of type "unprotected", shown in grey, denote street segments that are designed for motor vehicles and lack protected bicycle infrastructure. Links of type "protected", shown in black, denote protected bicycle infrastructure—either alongside a street segment or off-street. If a node has only one type of links adjacent to it, we call the node either a protected node, shown in black, or an unprotected node, shown in grey. If a node has both protected and unprotected links adjacent to it, we call it a contact node, shown in blue.

We then define a gap as a shortest path between two contact nodes that consists only of unprotected links. This definition is based on the rationale that a gap should be a continuous piece of "missing" protected infrastructure, and it should be as short as possible. An example of a gap following this basic definition is illustrated in red in Fig. 3.

For identifying gaps in our Copenhagen data set, we applied the Dijkstra all-pair shortest path algorithm to the entire street network with links weighted by length. From the set of paths obtained, we discarded all paths that do not meet our gap definition, i.e. the start and end nodes must be contact nodes and all links must be unprotected. In this way, 9924 unique gaps were identified in our Copenhagen data set.

Discarding parallel paths

Before proceeding to evaluating the benefits of closing gaps, we must ask whether our working definition of "gaps" will yield a meaningful set of potential "missing links", or whether we need to refine our approach. Indeed, applying our definition to Copenhagen's street network reveals the problem of *parallel paths* that needs to be accounted for. This problem comes from the naive application of the shortest path algorithm which does not account for the common occurrence of protected off-street bicycle tracks that run in parallel to car lanes. In these cases, the shortest path algorithm with links weighted by length chooses the slightly shorter car path over the slightly longer bicycle path (see Fig. 4), and therefore undesirably detects a gap located on the car lane despite a protected bicycle track running next to it. The parallel paths problem is a consequence of applying the shortest path algorithm to a relatively high-resolution network layer. However, lowering the resolution is not an option, because using map data with a high resolution of the street segments is necessary for identifying the gaps that we are looking for. This is a well-know problem in transportation network modeling: If a high-resolution layer is given as input, solving a routing problem at a lower resolution is a non-trivial task (Perrine, Khani, and Ruiz-Juri 2015; Zhu and Chiu 2015).

We therefore applied the following mitigation strategy for parallel paths: For each identified gap $\mathbf{g}$, we first computed the detour factor $D(\mathbf{g}) = \frac{d_{\text{prot}}(\mathbf{g})}{d_{\text{all}}(\mathbf{g})}$, where $d_{\text{prot}}(\mathbf{g})$ and $d_{\text{all}}(\mathbf{g})$ are the shortest network path distances on the network of protected bicycle infrastructure and on the entire street network, respectively. We then set a minimum detour value $D_{\text{min}} = 1.5$ and discarded all previously identified gaps that had $D(\mathbf{g}) < D_{\text{min}}$.

We arrived at the detour factor value of 1.5 by manually comparing the results of applying a cut-off value for gap rank and the declustering heuristic (see sections *Gap prioritization* and *Gap declustering* below) first to the list of gaps with $D(\mathbf{g}) \geq D_{\text{min}}$ and then to the list of gaps with $D(\mathbf{g}) < D_{\text{min}}$, for different values of $1 < D_{\text{min}} < 2$. Setting $D_{\text{min}} = 1.5$ yielded the fewest false positives and false negatives. For gaps with a detour factor of $D(\mathbf{g}) \geq 1.5$, there were only 10% of false positives, that is gaps with a detour factor of over 1.5 that turned out to be parallel paths and had to be excluded manually. For the gaps with a detour factor $D(\mathbf{g}) < 1.5$, we found three types of gaps: 1) an expected high percentage of 49% of parallel edges, 2) in 43% of cases a partial overlap with gaps of a higher detour factor and therefore no substantial loss of information when excluded, 3) only 8% of false negatives, that is actual gaps on the bicycle network. The chosen detour factor therefore presents a reasonable trade-off between minimizing false positives (roughly 10% of the gaps that had to be excluded manually) and loss of information (roughly 8% of automatically excluded gaps that were actually relevant). It is also in line with cyclist detour behavior reported in the literature (Reggiani et al. 2021).

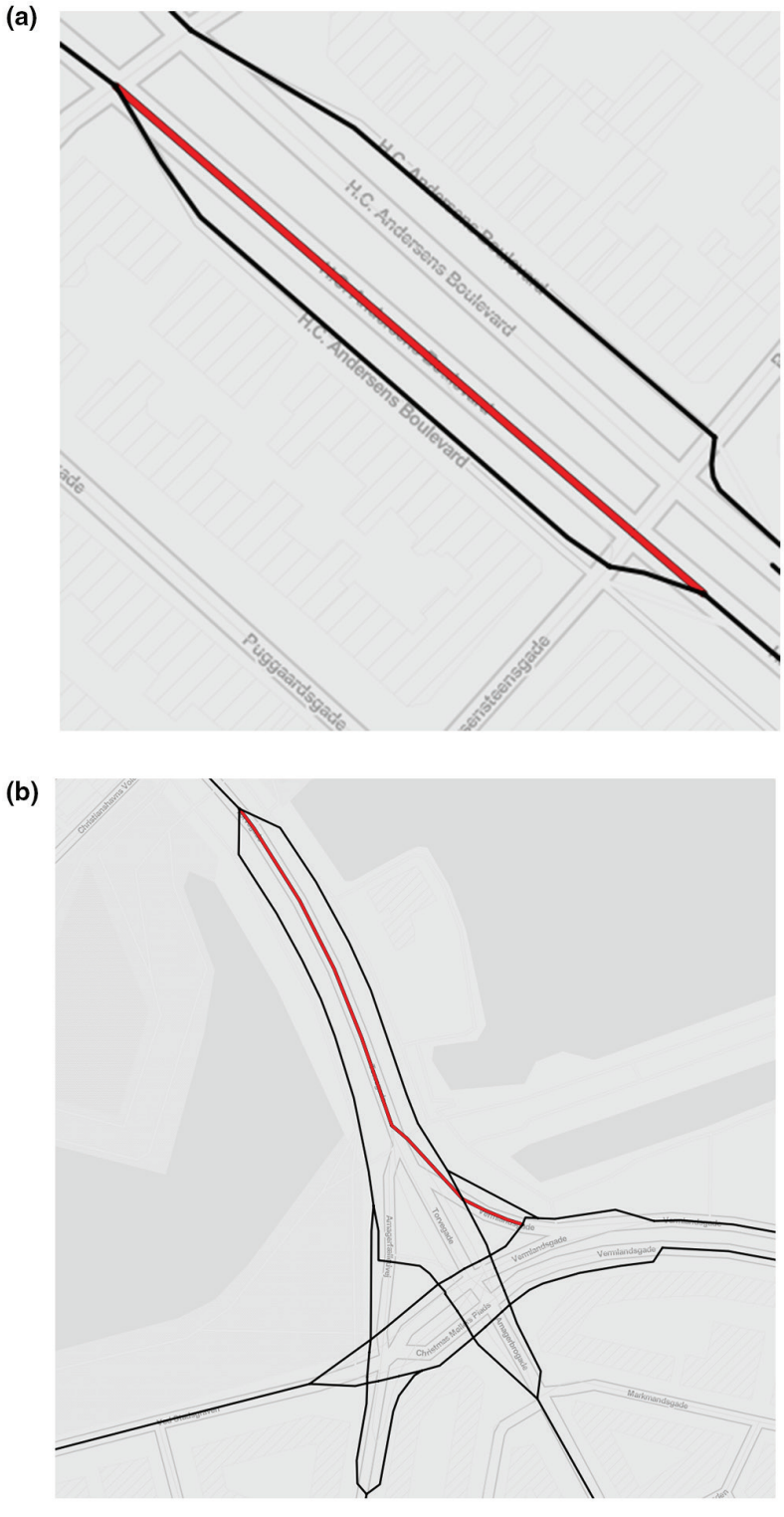


Figure 4. Two examples of parallel paths that a gap identification process must account for. The bicycle network is shown in black. The parallel paths along the car network are shown in red; they are only slightly longer and should not be identified as gaps. (a) Parallel path along H.C. Andersens Boulevard. (b) Parallel path along Torvegade. [Colour figure can be viewed at wileyonlinelibrary.com]

Excluding gaps with a detour factor below 1.5 from our analysis reduces the number of gaps from 9924 to 6603. This list of 6603 identified gaps is used as input for the next step of the IPDC procedure: gap prioritization.

Gap prioritization

Not all street segments that were identified so far as gaps are equally suitable for the construction of new bicycle infrastructure, nor are they equally relevant for the overall performance of the bicycle network. Consider and contrast two examples of locations without protected bicycle infrastructure: a residential street in a suburban area, versus a narrow bridge over a canal in the city center. To put a number on the priority of each of these gaps, we need to ask not only "How central is this missing link?" and "How much does it cost to close this gap?", but also "How many citizens will benefit from closing it?" Therefore, after having found all gaps which fit our topological definition, the next step is to evaluate the benefits of "closing a gap" (by installing protected bicycle infrastructure) for the overall performance of the bicycle network, and to prioritize the list of gaps by this benefit metric.

To quantify the benefit of "closing a gap", we start off with the rationale that the positive impact consists in reducing the number of meters that cyclists have to ride in the same space as motorized traffic. This is in line with the concept of "planning for the vulnerable", that is aiming to provide an inclusive transportation system by protecting the most vulnerable population groups—such as children, who ideally should never have to cycle in mixed traffic (McDonald 2012). If this concept was taken to the extreme, no single gap should be left unclosed, which is not a realistic goal. Therefore, we aim to approach this ideal *most effectively* by prioritizing gaps that lie on the most commonly taken bicycle routes. Using topological street network data only, the most common routes can be gauged quantitatively by selecting gaps with the highest link betweenness centrality weighted by gap length. Let us provide an example before the formal definition. Assume that gap A has a length of 10 m and a traffic volume of 50 cyclists in a time unit (e.g. during one hour); and gap B has a length of 20 m and a traffic volume of 15 cyclists. Then, by multiplying lengths with traffic volumes, we obtain the total number of meters cycled in mixed traffic: 500 m for gap A and 300 m for gap B. Closing gap A would avoid more meters cycled in mixed traffic, which is why gap A is ranked more relevant than gap B. In this case gap A is also shorter, therefore also more cost-efficient to close.

In order to apply this rationale, we estimated the number of cyclists on each link, i.e. the bicycle traffic flow through the network, based on the network topology, using betweenness centrality. Betweenness centrality, derived from an all-pair shortest path algorithm, is the most basic proxy for traffic demand. It assumes that for each possible origin-destination combination, there is one "cyclist unit" making their way through the network, always choosing the shortest possible path between origin and destination. Then the number of cyclists that use a specific link on their way through the network, divided by the total number of cyclists on the network, will yield the fraction of cyclists that we expect to find on this link. Thus, the betweenness centrality indicates how "central" or relevant a link is for the flow of cyclists through the whole network. Similar approaches based on betweenness centrality have previously been used to estimate bicycle and motorized traffic flow (McDaniel, Lowry, and Dixon 2014; Jayasinghe, Sano, and Nishiuchi 2015; Ye, Wu, and Fan 2016). This simple model can be refined arbitrarily by replacing betweenness with any other demand model, such as a gravity model, or with empirical flow data, but this refinement is outside the scope of this work.

There is a non-trivial dependence of flow-based centrality metrics on changes in network boundaries. This phenomenon, known as "network edge effect" (Okabe 2012) or "border effect" (Porta, Crucitti, and Latora 2006), also has relevant implications for equity considerations, since centrality metrics like betweenness have an inherent bias towards the center of the network. To account for this

network edge effect, we introduced a cut-off radius λ for the set of shortest paths, based on which the centrality metrics are computed (Gil 2017; Yamaoka, Kumakoshi, and Yoshimura 2021). Setting this locality parameter $\lambda = \infty$ in the shortest path algorithm would consider the entire street network for origin-destination pairs, finding gaps that are most relevant for the whole city and have a tendency to be located more centrally. By contrast, we used $\lambda = 2500$m to include only destination nodes that are within a maximum path length of 2500 meters from each origin node, finding gaps that are relevant for sub-city scale flows, e.g. on district or neighborhood scales. We chose this particular length as it roughly corresponds to the average diameter of the administrative districts of Copenhagen (Trap Danmark 2021), and lies in the range of 2–4.9 km, which is the most frequent bicycle trip length range within Copenhagen (Københavns Kommune, Teknik- og Miljøforvaltningen 2021). By using a finite λ in our calculations, we obtain several benefits: the bias towards the center of the network is decreased; the local importance of the identified gaps can be regulated; and, lastly, computation time is substantially reduced, which is particularly relevant for larger cities.

We applied the locality parameter λ to the set of all shortest paths $\mathcal{P}$ to compute the link betweenness centrality $c_\lambda(l) = \sum_{d(i,j)<\lambda} n_l(i,j)$ for each link l, where $n_l(i,j)$ is the number of times the link l appears in $\mathcal{P}$. By multiplying the betweenness centrality $c_\lambda(l)$ of link l by its length $L(l)$, we obtained the total number of expected meters cycled on this link, the link closure benefit $B^*_\lambda(l)$. Since a gap $\mathbf{g}$ can consist of several links, the gap closure benefit $B^*_\lambda(\mathbf{g})$ is obtained from adding up the link closure benefits of each of the links l:

$$B^*_\lambda(\mathbf{g}) = \sum_{l \in \mathbf{g}} c_\lambda(l) \cdot L(l) \tag{1}$$

As a last step, we account for cost-efficiency. We assume for simplicity, and in line with previous studies (Mauttone et al. 2017), that construction costs are generally proportional to facility length. We therefore divide the expected meters cycled $B^*_\lambda(\mathbf{g})$ by the gap's total length $L(\mathbf{g}) = \sum_{l \in \mathbf{g}} L(l)$ and thus obtain the expected meters cycled *per investment unit* that would be avoided if the gap was closed:

$$B_\lambda(\mathbf{g}) = \frac{B^*_\lambda(\mathbf{g})}{L(\mathbf{g})} \tag{2}$$

This model is extendable with further weights, for example with data on specific road hazards or stress levels (Furth, Mekuria, and Nixon 2016; Chen et al. 2017), or by a non-linear cost function. However, for sake of simplicity and generality, we do not assign any further weights here. This corresponds to the simplifying assumption that for each cyclist, every meter cycled jointly with motorized traffic equally contributes to the risk of getting injured or killed. Note that for the case of gaps that consist of only one link, equation (2) simplifies to $B_\lambda(\mathbf{g}) = c_\lambda(l)$ since the two expressions for total gap length cancel out. From here onwards, we drop the index λ for simplicity and denote the gap closure benefit as $B(\mathbf{g})$.

The benefit metric $B(\mathbf{g})$ will be used for gap prioritization. To summarize, it expresses the benefits of closing a gap $\mathbf{g}$ in terms of number of expected meters cycled in mixed traffic per unit of investment. Fig. 5 (left) shows the distribution of $B(\mathbf{g})$ for Copenhagen within the list of 6603 gaps that were found with a minimum detour of $D_{\min} \geq 1.5$. This distribution shows a large heterogeneity of benefits due to the heavy-tailed distribution of betweenness in street networks (Kirkley et al. 2018). In other words, there is a small subset of highest-ranked gaps which account for a substantial amount

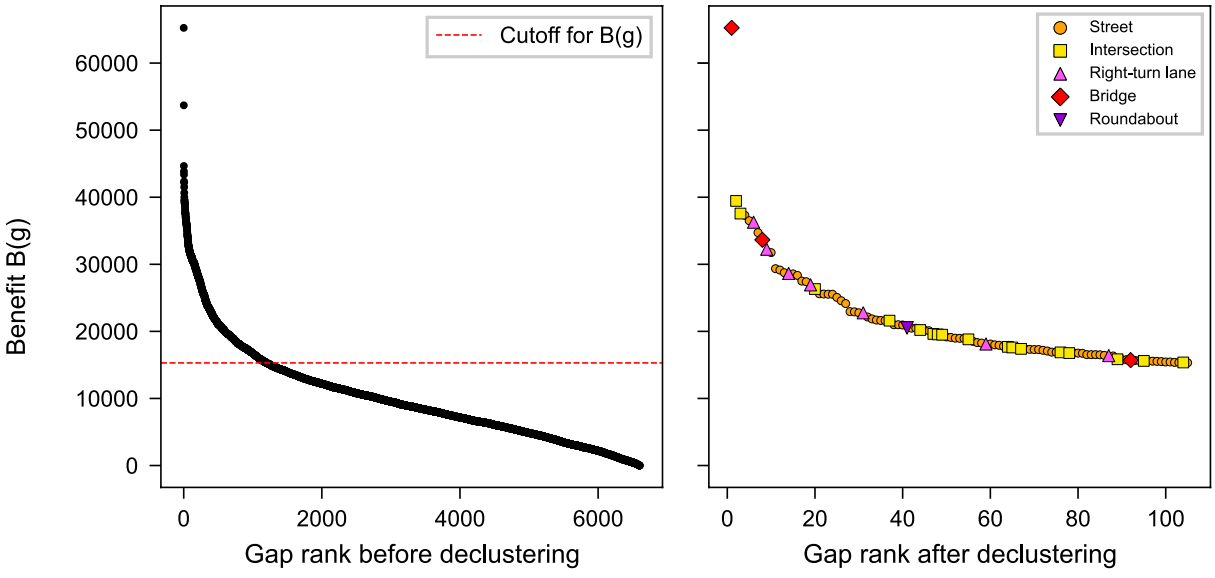


Figure 5. Heterogeneity of gap closure benefits. The distribution of the benefit metric $B(\mathbf{g})$ for the Copenhagen gaps before (left) and after (right) prioritization, declustering and error correction. The highest ranked gaps have a much higher benefit for the overall network; bridges tendentially fall into this category. The dashed red line on the left plot shows the benefit cut-off value used in the prioritization step. The colors in the left plot represent different gap classes (see section *Gap classification*). [Colour figure can be viewed at wileyonlinelibrary.com]

of the total benefit. After inspecting the heterogeneity of benefits, Fig. 5 (left), we chose a cut-off at $B(\mathbf{g}) \geq 15\,000$ where the growth of the rank-ordered benefits changes qualitatively, thereby selecting approximately the highest ranked 20% of gaps. This selection provides an ordered list of the 1199 highest-ranked gaps which is used as input to the next step of our IPDC procedure: gap declustering.

Gap declustering

In many cases, two or more prioritized gaps partially overlap, forming a *gap cluster* that is not a simple path anymore. An example of gap clustering in Copenhagen is shown in Fig. 6 (a): Because the intersection of C.F. Richs Vej and Grøndals Parkvej is represented by several network nodes, all shortest paths to destination node D from any of the origin nodes A, B, C are classified as gaps and display similar benefit values $B(\mathbf{g})$. This gap cluster example illustrates that it is not always meaningful to provide all street segments that constitute a gap cluster with protected bicycle infrastructure. The appearance of gap clusters in the results can also be understood by recalling our gap-finding procedure and network characteristics: First, all network are considered as equally likely origins or destinations; second, gaps consist of car links and start and end on a contact node; and third, the network is characterized by a high node density for example at intersections of streets with multiple lanes. Taken together, these three points help explain that gaps will often consist of a combination of high and low centrality links, and moreover, gaps will often partially overlap, meaning that the same street segment will appear in several gaps—for example, the network link on Brønshøjvej appears in more than 100 of the 6603 gaps found.

The urban planning task of identifying the exact subnetworks within these gap clusters for the construction of infrastructure to "close the gap" is beyond the scope of the present study. However, we developed a declustering heuristic, described in detail in Appendix A, which is a first approach to break down a gap cluster into separate components that are simple paths, based

Figure 6. Example of gap clusters, and of the declustering heuristic. (a) Shown in red, on C.F. Richs Vej. The three gaps AD, BD and CD overlap and have similar closure benefits $B(\mathbf{g})$. Therefore, the three gaps are merged into one gap cluster to be handled jointly. (b–d) A declustering heuristic can help deciding which parts to retain and which ones to discard. (b) The links of the shown gap cluster are colored by betweenness centrality; darker tones represent higher values. The black borders indicate the declustered gaps after (c) one and (d) two runs of the heuristic. [Colour figure can be viewed at wileyonlinelibrary.com]

on the same benefit metric derived $B(\mathbf{g})$ from betweenness centrality. See Fig. 6 (b–d) for a non-trivial gap cluster with links colored by edge betweenness centrality values and the resulting declustered gaps. After applying the gap declustering heuristic, using the list of 1199 highest-ranked gaps as input, we obtain a list of 134 gaps. This list is used as input for the next step of the IPDC procedure: gap classification.

Gap classification

The last step in the IPDC procedure is the classification of gaps. The gap classification scheme described in this section was developed through manual inspection and on-site visits of the gaps identified in Copenhagen, hence it might need to be adapted or extended for other urban contexts in future research. We identified the following gap classes: Street (ST); intersection (IS); right-turn lane (RT); bridge (BR); roundabout (RA); and error (ER). This classification scheme is meant to facilitate both the interpretation of results from bicycle network analysis and the decision-making within a subsequent planning process. In this section, we describe the general concept behind each of the gap classes before discussing the specific results for Copenhagen.

Street

The gap class *street* corresponds most intuitively to the idea of a "gap in the bicycle network", i.e. a generic street segment without protected bicycle infrastructure. We define as street gap all mixed-traffic street segments whose both ends connect to protected bicycle infrastructure and that do not correspond to any of the other gap classes (bridge, intersection, roundabout, right-turn lane, or error).

Intersection

Missing links without protected bicycle infrastructure found at crossings of two or more streets are classified as *intersection*. Given that a high proportion of traffic crashes occur at intersections, intersection design is crucial for cyclist safety (Thomas and DeRobertis 2013). By the very nature of an intersection, a potential for conflict between traffic participants cannot be brought to zero; however, it can be minimized with appropriate planning (de Groot 2016). Intersection design deserves to be considered a discipline of its own right, and different network analysis methods than the one used in this study might need to be applied to explicitly identify problematic intersections from a bicycle network planning perspective (Furth, Mekuria, and Nixon 2016).

In the present study, we do not model intersections separately, but rather identify them as gap class in the last step of the procedure. Due to the underlying data structure in OSM, intersections could only be identified as gaps within the IPDC procedure if they contained at least one link, rather than just nodes. Additionally, there is a lack of consistency in OSM tagging when it comes to the designation of specific intersection segments as "protected" or "unprotected". The caveats of this approach are addressed in more detail in Section *Scope and limitations* and in Appendix A.

Right-turn lane

We classify intersection approaches where the lane for right-turning cars merges with the adjacent cycle lane as *right-turn lane* gaps. In such cases, the bicycle path ceases to be part of the protected bicycle network as it approaches an intersection, and cyclists are forced to mix with motorized traffic—see Fig. 7 for an example. This type of intersection approach design is a common feature of Copenhagen's bicycle network (Vejdirektoratet 2017). The Danish Road Directorate argues in favour of an intersection approach design with shortened cycle tracks where cyclists and cars mix for right turns (Sørensen, Jensen, and Hansen 2020; Vejdirektoratet 2020), while current international best practice standards recommend intersections that protect and prioritize cyclists (Wagenbuur 2014; de Groot 2016; National Association of City Transportation Officials (NACTO) 2019), such as dedicated bicycle queue areas and corner wedges or islands.

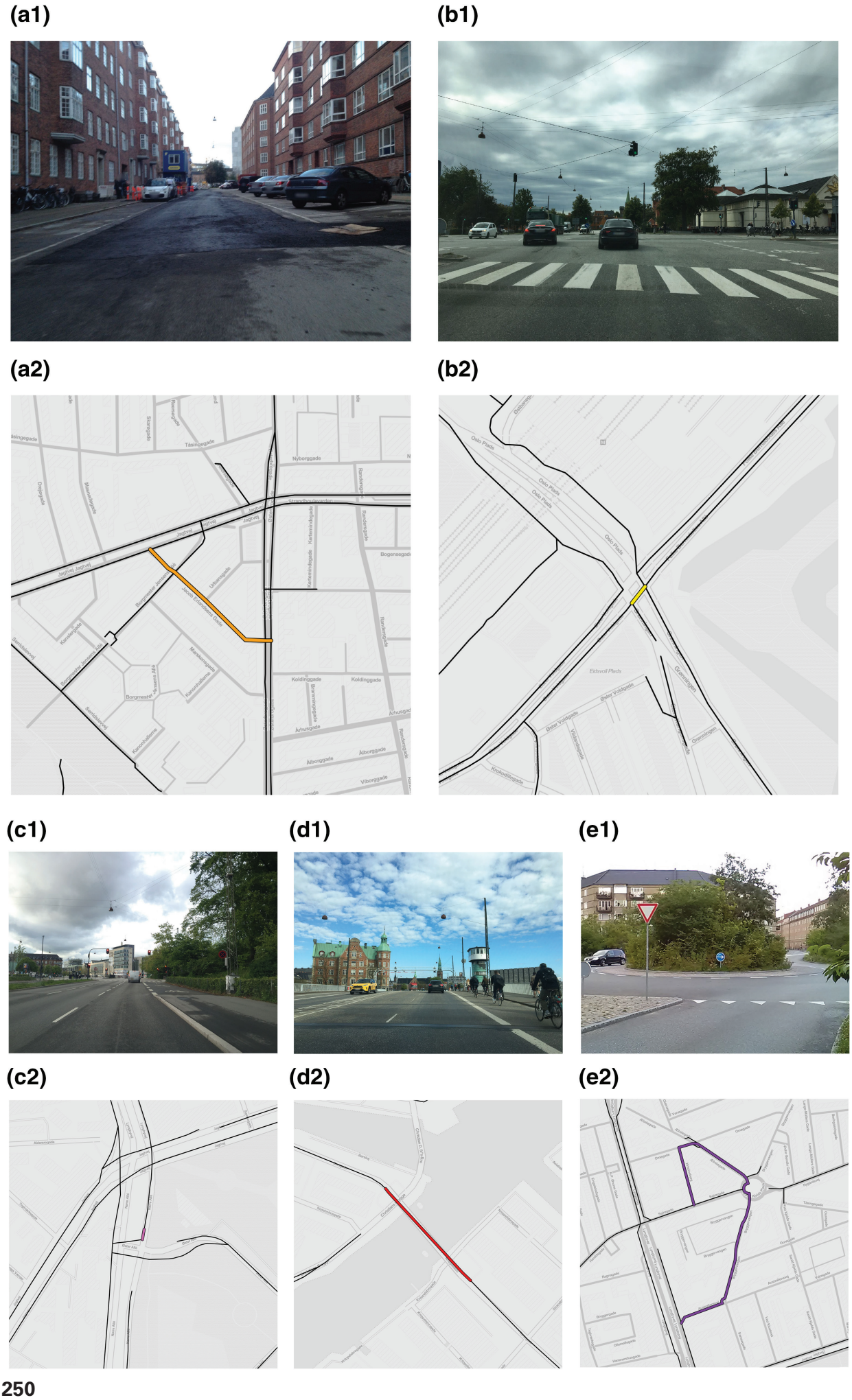


Figure 7. The five gap classes in Copenhagen. For each class, the highest ranked gap is shown. (a) Street gap (rank 4) on Jacob Erlandsens Gade, (b) Intersection gap (rank 2) at the intersection of Øster Vøldgade and Grønningen, (c) Right-turn lane gap (rank 6) at the right-turn from Nørre Allé to Øster Allé, (d) Bridge gap (rank 1) on Knippelsbro, (e) Roundabout gap (rank 41) at Sankt Kjelds Plads. [Colour figure can be viewed at wileyonlinelibrary.com]

Here we adhere to the international standards and to the rationale of demanding continuity for the network of protected bicycle infrastructure.

Bridge

We classify missing links on obstacle-crossing road segments as *bridge* gaps. In locations where there are physical barriers such as water bodies or railway tracks that have to be crossed, bridges play a particularly important function for connecting parts of the network and often constitute bottlenecks for traffic flow. At the same time, there are often inherent constraints to placing additional infrastructural elements on bridges due to limited physical space available (Wang et al. 2019).

Roundabout

Since requirements for roundabout design are not the same as for intersections, we separately define the gap class *roundabout*. Roundabouts are often considered to be the safer option for cyclists (Dufour 2010; Jensen 2017; U.S. Department of Transportation, Federal Highway Administration 2017), depending on traffic volume (de Groot 2016). A roundabout with more than one lane puts cyclists at danger (Dufour 2010). According to a recent literature review by Poudel and Singleton (2021), data from Northern Europe suggests that the number of bicycle crashes might actually be higher for roundabouts than for intersections. There are several roundabout design options focusing on cyclist safety (Sakshaug et al. 2010), such as the Zwolle roundabout, named after the Dutch city that first introduced it (Wagenbuur 2013; de Groot 2016).

Error

We classify gaps that have been identified by the IPDC procedure, but were not confirmed as such via visual inspection, as *errors*. There are two types of errors: *parallel paths* and *data issues*. Parallel paths, as described in section *Discarding parallel paths* above, are errors stemming from the routing problem in high resolution networks. Data issues are errors due to incorrect information on OSM. There are many possible reasons for errors in the OSM data: segments might be missing, mistagged, or outdated. The implications of OSM data quality on the results of this study are discussed in detail in section *Scope and limitations*.

Finding gaps in Copenhagen with IPDC: The top 105 gaps

In this section, we discuss the results of the IPDC procedure applied to the use case of Copenhagen. From the list of 134 gaps that were used as input for the last classification step of the IPDC procedure, we discarded 29 gaps classified as errors. We confirmed and classified the remaining gaps through manual inspection and on-site visits and obtained a list of 105 top priority gaps, which is the final result of the IPDC procedure applied to Copenhagen. The distribution of gap classes in the top 105 gaps is reported in Table 1 (class *error* excluded). The map in Fig. 8 gives

Table 1. Distribution of Gap Classes for the Top 105 Gaps in Copenhagen [Colour table can be viewed at wileyonlinelibrary.com]

Color	Acronym	Gap type	Count	Average benefit $\langle B \rangle_g$
	ST	Street	77	20,544
	IS	Intersection	17	20,925
	RT	Right-turn lane	7	25,911
	BR	Bridge	3	38,207
	RA	Roundabout	1	20,518

Bridges are most important.

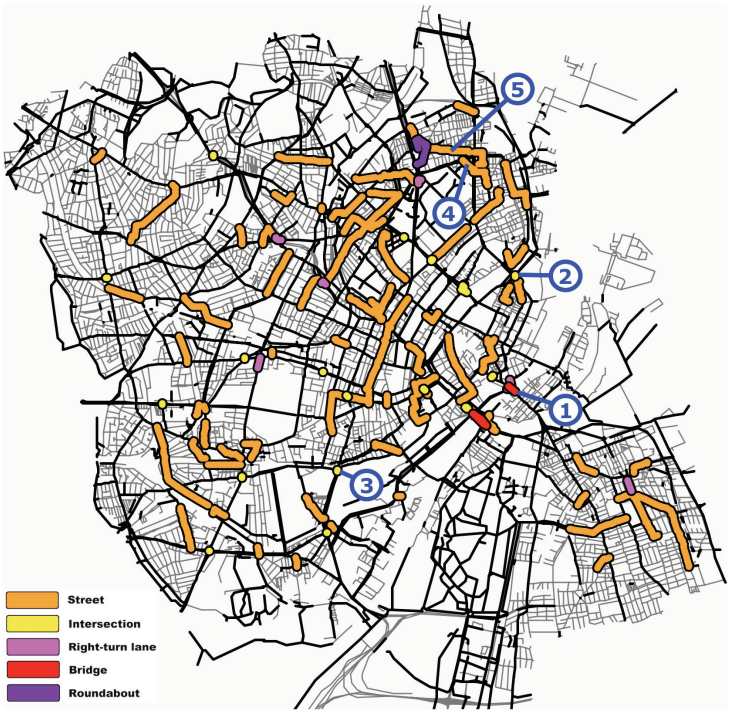


Figure 8. Overview map of top 105 gaps by class: streets in orange, intersections in yellow, bridges in red, right-turn lanes in pink, roundabouts in violet. Errors are shown in Appendix B. Numbered blue circles indicate the top 5 gaps (see Fig. 9 for detail plots). The street network is shown in grey, the bicycle network in black. See https://fixbike.net for an interactive version. [Colour figure can be viewed at wileyonlinelibrary.com]

an overview of all 105 gaps, with classes plotted by color. In the next sections, we summarize the results per gap class. A list of all 105 confirmed gaps with detail maps and addresses can be found in the SI and at https://fixbike.net/table.

Streets

Gaps classified as *street* constitute the majority of our final result (77 out of 105 gaps). Both visual analysis of the gap location and a comparison with Copenhagen's current Cycle Path Prioritization Plan (see section *Comparison with Copenhagen's Cycle Path Prioritization Plan* below) indicate that several of the identified street gaps might be confirmed as relevant by transport planning practitioners; for example, gap 5 on Tåsingegade (see Fig. 9), Gap 17 on Ålandsgade and Frankrigshusene, or gap 23 on Hamletsgade (see SI). Some of the identified street gaps are found on residential streets with presumably low traffic speed and volume,

so they would probably not be prioritized from a transport planning perspective in spite of their estimated local relevance indicated by high betweenness values. For example, gap 4 on Jacob Erlandsens Gade (see Fig. 7) shows that this short street—although low-traffic—is an important structural shortcut between the many paths connecting east of Østerbrogade and north of Jagtvej. For a refinement of the procedure, a further distinction of subcategories within the street gap class, both by road conditions (e.g. speed limit) and by empirical traffic volume data, if available, would be recommended, as it might help to estimate whether these links can be considered safely bikeable in spite of their lack of designated infrastructure (de Groot 2016).

Several street gaps come to lie within a locally sparse area of the network and are identified by the IPDC procedure due to the presence of small, isolated bicycle infrastructure elements in their vicinity. Examples are gap 46 on Valløvej and gap 62 on Oxford Allé (see SI). This is a direct consequence of our initial definition of a gap as a path between two bicycle infrastructure elements; hence, in network areas where no bicycle infrastructure at all is present, no gap will be identified, which makes the IPDC procedure less suitable for sparse network areas.

Intersections

With 17 out of the top 105 gaps, intersections are the second most common gap class. As outlined in the section on gap classification, due to both the data structure and data quality issues in OSM, the IPDC list of gaps classified as *intersection* should be understood as a non-exhaustive list of locations where checking for appropriate intersection design is recommended. It is noteworthy that most of the intersections identified by the IPDC procedure also received a considerable number of mentions as "busy intersections" in the citizen survey within the Cycle Plan. This is seen, for example, for gap 2 at Øster Voldgade and Grønningen and gap 3 at Enghave Vej and Vigerslev Allé (see Fig. 9), as well as gap 65 at H.C. Andersens Boulevard and Rystensteensgade (see SI).

Right-turn lanes

Seven out of the top 105 maps are classified as *right-turn lane*. Examples are gap 6 at Nørre Allé and Øster Allé (see Fig. 7), gap 9 at Backersvej and Øresundsvej and gap 14 at Borups Allé and Hillerodgade (see SI). The relatively low number of right-turn lanes in the top 105 gaps identified by the IPDC procedure can partially be explained by tagging inconsistency in OSM, already mentioned above with regard to intersections. We deem it likely that there is a significant number of false

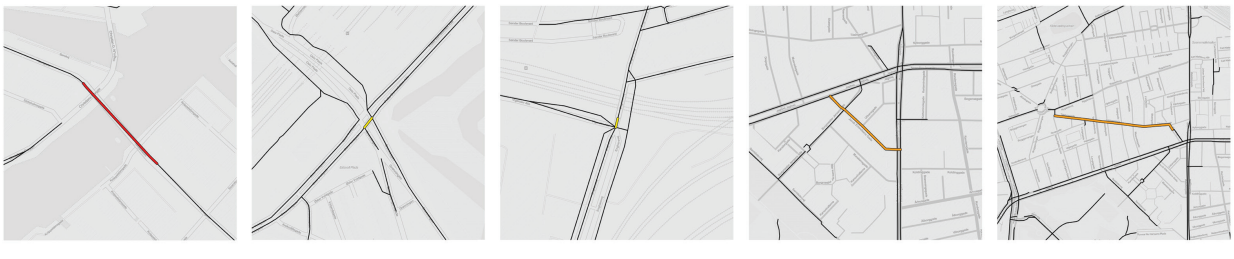


Figure 9. Detail plots of the top 5 gaps in Copenhagen. From left to right: Gap 1: Knippelsbro (bridge); Gap 2: Øster Voldgade and Sølvgade (intersection); Gap 3: Enghavevej and Vigerslev Allé (intersection); Gap 4: Jacob Erlandsens gade (street); Gap 5: Tåsingegade (street). All gaps can be explored at https://fixbike.net/table. [Colour figure can be viewed at wileyonlinelibrary.com]

negatives, i.e. right-turns that have not been identified as gaps by the IPDC procedure because they are tagged as "protected bicycle track" in OSM. Investigating both the OSM data quality and the objective and subjective safety implications of intersection approach design call for further research.

Bridges

There are 3 gaps classified as *bridge* within the top 105 gaps: gap 1 on Knippelsbro (see Fig. 7) and gaps 8 and 92 on Langebro (see SI). We have already argued for the physical separation of cyclists from motorized vehicles; it is of even higher relevance for the crossing of bridges (Melson, Duthie, and Boyles 2014). In the case of Copenhagen, bridges play a particularly relevant role as the city is situated on the two islands of Amager and Zealand, and harbours an extensive canal system. According to Copenhagen's latest Bicycle Account, 7 of the top 10 most heavily trafficked cycling stretches in the city are bridges (City of Copenhagen, Technical and Environmental Administration 2019). The first three stretches on that list are Dronning Louises Bro, Langebro and Knippelsbro. While Dronning Louises Bro is provided with protected bicycle infrastructure, Langebro and Knippelsbro are not. This aligns well with the results of the IPDC procedure, given that both Langebro and Knippelsbro are listed within the top 105 gaps. The Municipality of Copenhagen is currently in the process of upgrading the cycle lanes on both these bridges to cycle tracks (Københavns Kommune, Teknik- og Miljøforvaltningen 2017). The average benefit of closing a gap classified as *bridge* is almost twice as high as the average benefit for all other gap classes, see Table 1. This insight is in line with the underlying network topology—such "bridge edges" in infrastructure networks are important connections between otherwise separated or even disconnected parts of the network and therefore have particularly high betweenness centrality values.

Roundabouts

The only gap from the list of top 105 gaps classified as *roundabout* is gap 41 on Australiensvej/Bryggerivangen and Sankt Kjelds Plads. The gap contains two roundabouts: the bigger one, on Sankt Kjelds Plads, and the smaller one on the intersection of Australiensvej and Bryggervangen. The Sankt Kjelds Plads roundabout consists of only one lane where motorized vehicles and bicycles mix (see Fig. 7). Same as in the case of intersections, future work might consider the set of all roundabouts in the city of Copenhagen and examine their design from a cyclist safety perspective.

Errors

Out of the 29 gaps that have been discarded as errors within the last step of the IPDC procedure, classification, 15 were parallel paths (discussed in detail in section *Discarding parallel paths*) and 14 were data issues in OSM. Overview plots of all errors are found in *Appendix B*. Many of the parallel paths occur at large intersections or along streets with multiple lanes and bicycle infrastructure on both sides, for example, at Lyngbyvej or at the crossing of Frederikssundsvej and Borups Allé. Several parallel paths coincide with some of the busiest bicycle corridors in the city, such as Dybbølsbro and H.C. Andersens Boulevard, which is an encouraging observation for the use of betweenness centrality as a proxy for bicycle traffic flow. All data issues were due to missing tags for protected bicycle infrastructure in OSM, leading to the IPDC procedure identifying gaps in locations where protected bicycle infrastructure is already in place. While some of the missing OSM tags correspond to relatively

recent construction of infrastructure, others contain infrastructure that dates back more than a decade.

Comparison with Copenhagen's cycle path prioritization plan

The Municipality of Copenhagen's Technical and Environmental Administration (*Teknik- og Miljøforvaltningen*) regularly publishes a Cycle Path Prioritization Plan (*Cykelstiprioriteringsplan*, hereafter referred to as Cycle Plan). The current plan for the period 2017–2025 (Københavns Kommune, Teknik- og Miljøforvaltningen 2017) contains an overview of planned infrastructure improvements and measures targeted at increasing the modal share of cycling, split into five categories: new bicycle infrastructure (tracks, lanes, sharrows), improved intersection design, improvements of the Super Cycle Paths (*Supercykelstier*) network, improvements of the Green Cycle Routes (*Grønne cykelruter*) network, and finally, widening of existing cycle tracks. We conducted a comparative analysis of our list of top 105 prioritized gaps with the planned infrastructure improvements listed in the Cycle Plan across all categories except the last one (given that the width of bicycle infrastructure was not considered in this study). The comparison shows a considerable overlap, given that 46 out of 105 gaps identified by the IPDC procedure are found in locations that are also prioritized in the Cycle Plan. There is a particularly good overlap with the list of high priority routes for new cycle tracks (*højt prioriterede strækninger til nye cykelstier*): 19 out of 35 prioritized routes are included in our list of top 105 prioritized gaps (Københavns Kommune, Teknik- og Miljøforvaltningen 2017, p. 17). Further categories that show considerable overlap are the list of cycle lanes to be upgraded to cycle tracks (coinciding with 11 of our gaps); as well as identified missing links and planned upgrades of the Green Cycle Routes network (coinciding with 10 of our gaps).

The Cycle Plan also contains the results from a citizen survey on bicycle infrastructure improvements, conducted by the Municipality of Copenhagen in September and October 2016 (Københavns Kommune, Teknik- og Miljøforvaltningen 2017), which we used for a further qualitative assessment of the present study. Results from the citizen survey consist of a set of geocoded locations, indicated by respondents through clicking on a digital map, for each of the following categories: *Cykelsti mangler* (cycle track missing), *Cykelsti for smal* (cycle track too narrow) and *Kryds med stor traengsel* (busy intersection). We did not utilize responses from the category on too narrow cycle tracks, given that street width was not accounted for in the present study. Figure 10 provides an overview of the processed data from the citizen survey.

A qualitative comparison of our list of top 105 prioritized gaps with the citizen survey results shows considerable overlaps at several locations. Examples are shown in Fig. 11. In total, 71 out of our 105 gaps have at least one mention in the considered categories of the citizen survey. Although these overlaps are encouraging at first glance, there is a relevant caveat to consider. While participatory approaches can improve the equity impact of transportation plans (Boisjoly and Yengoh 2017), a failure to adequately design them might introduce biases and undermine the applicability of the findings (Schonlau et al. 2009; Nohr and Liew 2018). A reliable survey design should account for several bias/equity considerations, such as survey language, medium used, distribution channels and socio-demographic variables of respondents. We have no information about such considerations (or the lack thereof) for the survey data at hand. Therefore, if a location has no mentions in the citizen survey, it cannot be concluded that the infrastructure is already satisfactory there—it might be due to an undersampling of residents from that area.

Figure 10. Overview map of citizen survey data. Citizen responses on missing bicycle tracks and busy intersections are represented by blue dots. The street network is shown in grey, the bicycle network in black. [Colour figure can be viewed at wileyonlinelibrary.com]

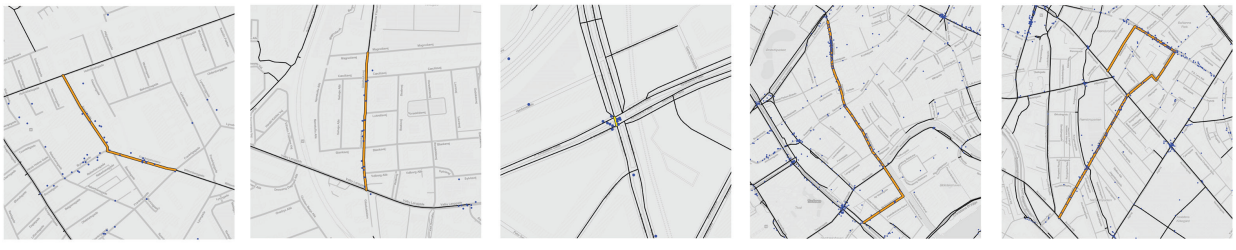


Figure 11. Five examples of overlaps between gaps found by the IPDC procedure and citizen survey results (blue dots). From left to right: Gap 17 on Ålandsgade and Frankrigshusene; Gap 30 on Gåsebæksvej; Gap 47 on the intersection of Enghavevej with P. Knudsens Gade; Gap 69 on Nørregade/Rådhusstræde/Ny Kongensgade; Gap 90 on Stefansgade/Gormsgade/Mimersgade. [Colour figure can be viewed at wileyonlinelibrary.com]

As long as such considerations in the citizen survey design are unclear, its results should not be regarded as reliable ground truth.

The partial overlap of our top 105 gaps with the locations prioritized by the Municipality of Copenhagen in the Cycle Plan, as well as with the citizen survey results, is a first proof of concept for the IPDC procedure. At the same time, the gaps from our results that do *not* show up in the Cycle Plan are of particular interest for further evaluation, enhancement of methods and decision-making. In a future dialogue with the Municipality of Copenhagen, the results from the IPDC procedure could be scrutinized to find out which gaps are actual missing links in the bicycle network of Copenhagen and possibly will be prioritized in future infrastructure investments; which gaps are less relevant from an urban planning perspective and indicate a necessity to adjust our method (e.g. by adding information on street type or non-protected bicycle

infrastructure to the analysis); and finally, which gaps have been wrongly identified due to data issues in OpenStreetMap. Thus, the comparison with Copenhagen's Cycle Plan demonstrate that the IPDC procedure has the potential to be used as automatized assistance tool and to successfully complement manual planning processes, while its results can and must be further scrutinized by urban planners.

Discussion

Our results from the Copenhagen case study suggest that benefits from bicycle infrastructure improvements for overall network quality are highly variable and location-dependent, with the gap class *bridge* showing the highest average benefits. We also find that network edge effects in transport network analysis might have detrimental implications for the population in the urban periphery, and that future work is needed to mitigate this bias. These findings illustrate the advantages of considering the network as a whole in the analysis, operating on the "macro level". In practice, bicycle infrastructure planning often is highly localized and guided by manual decision-making, taking place on a "micro level". Our results show that these two approaches should not be seen in competition, but rather as complementary to each other. This is illustrated by the application of the IPDC procedure to bicycle network of Copenhagen and the comparison of the findings with the city's Cycle Path Prioritization Plan, since potentially relevant results were obtained in spite of minimum data requirements. We are therefore optimistic about the potential of a computational, data-driven macro level approach to decision-making support for bicycle network planning. The IPDC procedure presented in this study is, however, just a first step towards this goal; in the following sections, we discuss the scope and limitations of our approach, as well as further work needed.

Scope and limitations

The potentially most substantial limitation for the results of this study is data quality in OpenStreetMap. OSM data is crowdsourced, which allows for the integration of local knowledge and the provision of open source data, but at the same time often leads to data quality issues due to different skill levels within the mapping community and a lack of coherence in tag criteria applications (Kaur and Singh 2018). In addition, OSM data quality significantly varies by location (Mooney, Corcoran, and Winstanley 2010; Haklay 2010). A broader quantitative assessment of OSM data quality is still an open research question (Jacobs and Mitchell 2020; Yeboah et al. 2021), particularly for bicycle infrastructure (Ferster et al. 2020). In our results from the IPDC procedure, many of the identified gaps which were discarded as data issues were due to outdated OSM tags, with substantial portions of recently built bicycle infrastructure not yet included in the OSM data. While the number of tagging edits might potentially be used as a workaround for estimating whether the tag is up-to-date (Line 2021), ideally the implementation of new bicycle infrastructure elements would go hand in hand with the corresponding update in OSM. Another issue is the lack of coherence in bicycle infrastructure tagging. For example, right-turn lanes where the bicycle track merges with a car lane are sometimes marked as protected bicycle infrastructure; the same goes for unprotected intersections which separate two stretches of protected bicycle infrastructure. Therefore, the definitions of bicycle infrastructure categories within OSM (OpenStreetMap Contributors 2021) might be scrutinized from the viewpoint of intelligibility in order to enhance correct

and coherent identification of bicycle infrastructure by mappers across differing local contexts (Ferster et al. 2020).

A further limitation consists in our simplified conceptualization of street and bicycle networks based on protected bicycle infrastructure availability. Our study considers only protected bicycle infrastructure as part of the bicycle network, but unprotected bicycle infrastructure can also be an adequate design solution under certain conditions (de Groot 2016). By binarizing street categories into "protected" and "unprotected", we assume that it is equally undesirable to cycle on any of the car-only streets, whereas in reality the propensity to cycle in mixed traffic highly depends on such factors as road type, traffic flow, and number of lanes. While the IPDC procedure for Copenhagen delivered relevant findings in spite of these simplifications, results could be further scrutinized by enhancing the network model through a more fine-grained differentiation of road and bicycle infrastructure types. The level of detail that can be introduced into the network model will depend on the level of data availability.

Similarly, both the calculation of the benefit and the estimation of bicycle traffic flow within the IPDC procedure could be enhanced in case corresponding data is available. In this study, we assumed minimum data availability and estimated traffic flows and construction costs based only on topological network properties. This assumption was followed intentionally, since our aim was to develop a general method. However, if empirical traffic flow measurements, origin-destination tables, census data etc. are available (Olmos et al. 2020), the calculation of a flow centrality, construction costs and total benefit for each network element could be made more accurate. A related caveat in relation to betweenness centrality is the finite λ parameter that we introduced, as a first attempt to partially mitigate the bias towards the network center. The equity implications of centrality metrics have only recently started to be discussed and there is a knowledge gap regarding their quantification (Jafino, Kwakkel, and Verbraeck 2020; Jafino 2021; Yamaoka, Kumakoshi, and Yoshimura 2021). A systematic analysis of such network edge effects would therefore be of high relevance. However, it also goes beyond the scope of the present study, so future work in that regard is urgently called for.

A further simplification is the assumption that cyclists always choose the shortest path from A to B. This is implied in our definition of betweenness centrality, since the shortest path computations on the network are performed with link weight set equal to link length. Several previous studies have accounted for cyclist preferences in shortest paths computations by providing links with a weighting factor that is based on additional features which quantify link attractiveness for cyclists (Broach, Dill, and Gliebe 2012; Furth, Mekuria, and Nixon 2016; Cervero, Denman, and Jin 2019; Boisjoly, Lachapelle, and El-Geneidy 2020). While such an approach may result in more realistic cyclist flow estimations and mitigates the parallel paths problem for some locations on the network, it comes at a considerable cost: A feature-based weighting factor for network links is highly context-dependent, based on potentially subjective cyclist preferences, and constitutes an additional parameter with a non-trivial impact on the shortest path calculations. We therefore explicitly decided not to consider link weighting factors other than link length for our shortest path computations, but extending our model in this respect would be straightforward.

Lastly, given that the classification scheme presented in this study was derived from a qualitative analysis of results for Copenhagen, it might need to be modified for other local contexts. The same goes for the parameters $D_{\min}$ (minimum detour) and $B(\mathbf{g})$ (cut-off benefit) which have been selected for Copenhagen. Appropriate values for both parameters have been derived empirically, but no statement can be made concerning appropriate parameter values for other cities. Although these

two thresholds were selected manually, we do not expect this to affect the robustness of our results—rather, we expect an adjustment of thresholds to mostly impact the number of gaps found.

Future research

Based on the findings from this study, we anticipate four major lines of future research. First, the IPDC procedure presented here should be further improved. As discussed in the previous section, *Scope and limitations*, there are numerous ways to make the IPDC procedure more accurate, including a testing of its applicability to other locations. Second, we call for the urgent development of a solid computational basis for data-driven bicycle network planning, following recent first steps (Olmos et al. 2020; Natera Orozco et al. 2020a; Mahfouz, Arcaute, and Lovelace 2021). We deem it particularly relevant to consider multimodality and the multiplex transport network of a city as a whole (Natera Orozco et al. 2021), and to include equity considerations as an integral part of the network analysis process (Gössling 2016; Pereira, Schwanen, and Banister 2017; Jafino 2021). Third, we emphasize the importance of bicycle infrastructure data quality, availability, and coherence (Ferster et al. 2020). Access to high quality data is a necessary precondition to provide a scientific basis for any substantial systemic shift towards more active mobility. Fourth and lastly, in line with our call for better cycling data and for data-driven planning approaches, we recommend to account for limited data availability and corresponding mitigation options in any future work on bicycle network planning.

Conclusion

In this study, we developed the IPDC procedure for identifying, prioritizing, clustering and classifying gaps in urban bicycle networks. Our method is based only on topological network properties and thus has minimal data requirements. We applied the IPDC procedure to the city of Copenhagen and obtained a list of 105 top priority gaps. A comparison of our results with the city's most recent Cycle Path Prioritization Plan showed substantial overlaps, both with citizen input on missing bicycle network links and with the city's list of prioritized locations for the construction of new bicycle infrastructure. The IPDC procedure demonstrates how data-driven network analysis on a city-wide scale can meaningfully complement manual planning processes. We therefore consider this study a further crucial step towards a consolidation of computational methods for bicycle network analysis.

Acknowledgements

The authors would like to thank Ahmed El-Geneidy, Ane Rahbek Vierø, Kim Sneppen, Marie Kåstrup and Robin Lovelace for their valuable input and comments, and Københavns Kommune for providing us with the Cycle Path Prioritization Plan data. We gratefully acknowledge the open source data and software that this article is based on: Map data copyrighted by OpenStreetMap contributors and available from https://www.openstreetmap.org; map tiles by Stamen Design, under CC BY 3.0; and images from Mapillary licensed under Creative Commons Share Alike (CC BY-SA) 4.0.

Data Availability Statement

The code for the IPDC procedure, as well as the OSM data used as input for the Copenhagen case study, is available on GitHub: https://github.com/anastassiavybornova/bikenwgaps.

APPENDIX A

Data acquisition and processing

We describe the details of data acquisition and processing, as carried out in the present study, within the following subsections: Data source and data structure; Simplification of OSM data; Representation of intersections in OSM data; and lastly, Declustering. All code described in this section is available on GitHub: https://github.com/anastassiavybornova/bikenwgaps.

Data source and data structure

For data acquisition and data processing we used Python and OSMnx (Boeing 2017). The main data source is OpenStreetMap. The input for the case study on Copenhagen consists of GIS vector data of geographic objects which together form the street network of Copenhagen (streets and intersections, bridges, roundabouts, parking lots, paths through green areas etc.) Intersections are represented as points with geographic coordinates, and street segments are represented as sequences of points. In our network derived from the data, intersections of street segments are interpreted as network nodes, and street segments are interpreted as links.

All input data was downloaded from OSM in February 2021 in csv file format. Data sets were acquired separately for two partially overlapping networks, which, when combined, form the street network of the municipalities of Copenhagen and Frederiksberg: the network of car infrastructure and the network of protected bicycle infrastructure, or, in more simple terms, the car network and the bicycle network. The limits of the two networks coincide with municipality boundaries, which introduces a cut into the continuous fabric of the street network of the Greater Copenhagen area (see the discussion on network edge effects in the section *Gap prioritization*). For each of the two networks, two data sets were generated through OSMnx: one for the nodes and one for the links. Each node from the data set has the attributes geocoordinates and OSM ID; each link has the attributes geocoordinates, OSM ID, length, street name and oneway/twoway indication, as well as several attributes which have not been used within the scope of this study, such as type of highway and speed limit.

The data on car and bicycle nodes was combined into one data set and the parameter "node type" was added. Nodes that appeared only in the bicycle data set were assigned the type "protected" and nodes that appeared only in the car data set were assigned the type "unprotected". Nodes that appeared in both data sets were assigned the type "contact". After this, duplicates were removed. The same procedure was applied to the car and bicycle link data sets: they were merged into one data set with the "link type" parameter set to "unprotected" (if the link appeared only in the car data set) or to "protected" (if the link appeared in the bicycle data set or in both data sets). Duplicated links, that is links with same length and type but opposite origin/destination nodes, were removed.

A graph object was created from the resulting data set using the Python's networkx library. The resulting network had 77 disconnected components, out of which only the largest connected component was kept, while all other disconnected components were dismissed as negligible for the sake of simplicity. In the real street network of the city, disconnected components, that is street segments that are not accessible from any other street segment, are quite rare. The appearance of disconnected components in our data set is mostly due to data quality issues, for example missing street segments that should have been classified as protected links.

Simplification of OSM data

Within OSM data, prior to further processing, a curved street is represented by a sequence of several points in geocoordinates, which are connected by straight lines. We shall call the corresponding degree-two nodes, which are introduced only for the sake of preserving the physical shape of a link, "auxiliary". The presence of auxiliary nodes in the data set strongly biases the degree distribution of the network towards $d = 2$. The network can be simplified by replacing a sequence of straight links and their corresponding auxiliary nodes by a single polygon link, while preserving the data on length and coordinates of the aggregated links. OSMnx has a built-in function to export already simplified data sets. For our purposes, however, the simplification had to be carried out on the combined network of protected and unprotected links (as opposed to separately simplifying the car and bicycle network, which is an already automated functionality in OSMnx). This is because nodes which are auxiliary in only one of the two networks would otherwise disappear from the data set, and information on connections and partial overlaps between the car and bicycle networks would be lost in case of separate simplification. Therefore, a

network was created from the merged data set of protected/unprotected links and protected/unprotected/contact nodes. Then, a simplification algorithm, described in Box 1, was applied to the network to remove all auxiliary nodes.

For the data set used in the present study, the simplification algorithm terminates after seven runs; the highest number of auxiliary nodes associated with a link in the final, simplified network is 54. The only degree-two nodes that appear in the data set after simplification are either meeting points of two links of different types or nodes that are kept to represent loops on the network while maintaining the network simple, i.e. without parallel links. As expected, the degree distribution of the simplified network significantly differs from the original one, shifting from a high to a low percentage of degree-two nodes.

Input: Network H with auxiliary nodes
Output: Network H' without auxiliary nodes

```
while auxiliary nodes in H do
    for node in H do
        if node degree d(n) = 2 and links incident on node have the same type then
            place node in stack
        end
    end
    while stack is not empty do
        take random node n from stack;
        if neighbours of n are neighbours themselves then
            remove node n from stack;
        else
            remove two links incident on node n from link set of network H;
            add new link connecting two neighbours of n to the link set of network H;
            set length attribute of new link to sum of lengths of removed links;
            add geocoordinates of removed links to geocoordinate attribute of new link;
            remove node n from node set of network H;
            remove node n and, if applicable, its two neighbours from stack;
        end
    end
end
```

The final outcome of the data preprocessing is the car and bicycle network of Copenhagen, represented by a simple, loop-free, undirected graph with no auxiliary nodes, where each link has two attributes: *type* ("protected" or "unprotected") and *length*, and each node has the attribute *type* ("protected", "unprotected", or "contact").ted", or "contact").

Representation of intersections in OSM data

Within the OSM data structure, intersections of smaller spatial extent appear as single nodes (a node representing the crossing of two streets), while larger ones appear as a set of nodes and links (each node representing the intersect of two or more lanes—see the example in Fig. 6 where nodes A, B and C are all part of the same intersection). As a rule, but not exclusively, this is the case when at least one of the intersecting streets is bidirectional. Keeping this representation of larger intersections within the data structure allows for the identification of unprotected crossings, that lie on an otherwise protected bicycle track, as gaps in the bicycle network. However, this method of identifying unprotected crossings is by far not exhaustive. This has several reasons. First, due to the data structure, the IPDC procedure does not recognize unprotected intersections that are represented by single nodes in the network model as gaps. Second, even with a clearly outlined set of intersection design criteria at hand which would enable us to discard protected intersections

from the gap list, the incoherence of intersection tagging in OSM results in numerous false negatives and false positives: intersections with a protected crossing for cyclists are often tagged as unprotected bicycle infrastructure; intersections without any bicycle infrastructure are often tagged as part of the cycle track they are actually interrupting.

Declustering

To decluster the partially overlapping gaps identified by the IPDC procedure, we developed a simple declustering heuristic, which is described in Box 2. Within the declustering process, gaps with a benefit metric of at least $B(\mathbf{g})_{\min}$ are combined into a network C, after which each disconnected component of C is declustered separately. Declustered gaps are added to the declustered gap list d. Gaps that obtain a benefit metric below $B(\mathbf{g})_{\min}$ are discarded. The list of remaining gaps is the output of the declustering heuristic and the input for the next step in the IPDC procedure (classification).

In the present study, the benefit metric of $B(\mathbf{g})_{\min} = 15\,000$ is used as cut-off value, which results in a list of 1199 gaps as input for the declustering heuristic. The gap network C consists of 101 disconnected components (gap clusters). The resulting gap set contains 168 declustered gaps, out of which 34 are discarded due to their lower values of $B(\mathbf{g}) < B(\mathbf{g})_{\min}$; the final output d is a list of 134 gaps.

Input: List c of partially overlapping gaps; cut-off benefit metric $B(\mathbf{g})_{\min}$
Output: List d of non-overlapping gaps

Remove gaps with $B(\mathbf{g}) < (\mathbf{g})_{\min}$ from c ;
Combine gaps c into network C ;
Decompose network C into a list of disconnected components dc ;
for *comp in dc* **do**
  while *comp is not empty* **do**
    compute all shortest paths between nodes $n \in comp | d(n) \neq 2$;
    compute benefit metric $B(\mathbf{g})$ for each path ;
    find path $p_{\max}$ with highest value $B(\mathbf{g})_{\max}$;
    add $p_{\max}$ to final gap list d ;
    remove $p_{\max}$ from $comp$;
  end
end
Remove gaps with $B(\mathbf{g}) < (\mathbf{g})_{\min}$ from d

APPENDIX B

Error plot

See Fig. B1.

Figure B1. Errors in the list of top ranked gaps in Copenhagen. 29 out of 134 gaps identified by the IPDC procedure in the Copenhagen network that have been discarded as errors: data issues in light blue; parallel paths in light green. [Colour figure can be viewed at wileyonlinelibrary.com]

References

Abad, L., and L. Van der Meer (2018). "Quantifying Bicycle Network Connectivity in Lisbon Using Open Data." *Information* 9(11), 287.

Battiston, F., V. Nicosia, and V. Latora (2014). "Structural Measures for Multiplex Networks." *Physical Review E* 89, 032804.

Batty, M. (2013). The New Science of Cities. Cambridge, MA, London, UK: MIT Press.

Boeing, G. (2017). "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." *Computers, Environment and Urban Systems* 65, 126–39.

Boisjoly, G., U. Lachapelle, and A. El-Geneidy (2020). "Bicycle Network Performance: Assessing the Directness of Bicycle Facilities Through Connectivity Measures, a Montreal, Canada Case Study." *International Journal of Sustainable Transportation* 14(8), 620–34.

Boisjoly, G., and G. T. Yengoh (2017). "Opening the Door to Social Equity: Local and Participatory Approaches to Transportation Planning in Montreal." *European Transport Research Review* 9(3), 1–21.

Broach, J., J. Dill, and J. Gliebe (2012). "Where Do Cyclists Ride? A Route Choice Model Developed with Revealed Preference GPS Data." *Transportation Research Part A: Policy and Practice* 46(10), 1730–40.

Buehler, R., and J. Dill (2016). "Bikeway Networks: A Review of Effects on Cycling." *Transport Reviews* 36(1), 9–27.

Carstensen, T. A., A. S. Olafsson, N. M. Bech, T. S. Poulsen, and C. Zhao (2015). "The Spatio-Temporal Development of Copenhagen's Bicycle Infrastructure 1912–2013." *Geografisk Tidsskrift-Danish Journal of Geography* 115(2), 142–56.

Cervero, R., S. Denman, and Y. Jin (2019). "Network Design, Built and Natural Environments, and Bicycle Commuting: Evidence from British Cities and Towns." *Transport Policy* 74, 153–64.

Chen, C., J. C. Anderson, H. Wang, Y. Wang, R. Vogt, and S. Hernandez (2017). "How Bicycle Level of Traffic Stress Correlate with Reported Cyclist Accidents Injury Severities: A Geospatial and Mixed Logit Analysis." *Accident Analysis & Prevention* 108, 234–44.

City of Copenhagen, Technical and Environmental Administration (2019). "The Bicycle Account 2018." Copenhagen City of Cyclists.

Cooper, C. H. V. (2018). "Predictive Spatial Network Analysis for High-Resolution Transport Modeling, Applied to Cyclist Flows, Mode Choice, and Targeting Investment." *International Journal of Sustainable Transportation* 12(10), 714–24.

de Groot, R. (2016). Design Manual for Bicycle Traffic. Ede, the Netherlands: CROW.

Dill, J., and J. Gliebe (2008). Understanding and Measuring Bicycling Behavior: A Focus on Travel Time and Route Choice. https://pdxscholar.library.pdx.edu/usp_fac/28.

Dufour, D. (2010). "PRESTO Cycling Policy Guide." *Cycling Infrastructure*. https://ec.europa.eu/transport/sites/default/files/cycling-guidance/presto_cycling_policy_guide_infrastructure.pdf

Feddes, F., M. de Lange, and M. te Brömmelstroet (2020). "Hard Work in Paradise. The Contested Making of Amsterdam as a Cycling City." The Politics of Cycling Infrastructure: Spaces and (In) Equality. Bristol: Policy Press.

Ferster, C., J. Fischer, K. Manaugh, T. Nelson, and M. Winters (2020). "Using OpenStreetMap to Inventory Bicycle Infrastructure: A Comparison with Open Data from Cities." *International Journal of Sustainable Transportation* 14(1), 64–73.

Furth, P. G., M. C. Mekuria, and H. Nixon (2016). "Network Connectivity for Low-Stress Bicycling." *Transportation Research Record* 2587(1), 41–9.

Gil, J. (2017). "Street Network Analysis "Edge Effects": Examining the Sensitivity of Centrality Measures to Boundary Conditions." *Environment and Planning B: Urban Analytics and City Science* 44(5), 819–36.

Gössling, S. (2016). "Urban Transport Justice." *Journal of Transport Geography* 54, 1–9.

Gössling, S., A. Choi, K. Dekker, and D. Metzler (2019). "The Social Cost of Automobility, Cycling and Walking in the European Union." *Ecological Economics* 158, 65–74.

Haklay, M. (2010). "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B: Planning and Design* 37(4), 682–703.

High-level Advisory Group on Sustainable Transport (2016). "Mobilizing Sustainable Transport for Development." Analysis and Policy Recommendations from the United Nations Secretary-General's High-Level Advisory Group on Sustainable Transport. https://sdgs.un.org/publications/mobilizing-sustainable-transport-development-18045.

Jacobs, K. T., and S. W. Mitchell (2020). "OpenStreetMap Quality Assessment Using Unsupervised Machine Learning Methods." *Transactions in GIS* 24(5), 1280–98.

Jafino, B. A. (2021). "An Equity-Based Transport Network Criticality Analysis." *Transportation Research Part A: Policy and Practice* 144, 204–21.

Jafino, B. A., J. Kwakkel, and A. Verbraeck (2020). "Transport Network Criticality Metrics: A Comparative Analysis and a Guideline for Selection." *Transport Reviews* 40(2), 241–64.

Jayasinghe, A., K. Sano, and H. Nishiuchi (2015). "Explaining Traffic Flow Patterns Using Centrality Measures." *International Journal for Traffic and Transport Engineering* 5(2), 134–49.

Jensen, S. U. (2017). "Safe Roundabouts for Cyclists." *Accident Analysis & Prevention* 105, 30–37.

Kaur, J., and J. Singh (2018). "An Automated Approach for Quality Assessment of OpenStreetMap Data." In 2018 International Conference on Computing, Power and Communication Technologies (GUCON). 707–12.

Københavns Kommune, Teknik- og Miljøforvaltningen (2017). Cykelsti-Prioriteringsplan 2017–2025. https://byudvikling.kk.dk/sites/byudvikling.kk.dk/files/cykelstiprioriteringsplan-2017-2025pdf-_1620.pdf.

Københavns Kommune, Teknik- og Miljøforvaltningen (2021). "Fra god til verdens beste." Københavns cykelstrategi 2011–2025. https://kk.sites.itera.dk/apps/kk_pub2/?mode=detalje&id=818.

Kirkley, A., H. Barbosa, M. Barthelemy, and G. Ghoshal (2018). "From the Betweenness Centrality in Street Networks to Structural Invariants in Random Planar Graphs." *Nature Communications* 9(1), 2501.

Koglin, T., and T. Rye (2014). "The Marginalisation of Bicycling in Modernist Urban Transport Planning." *Journal of Transport & Health* 1(4), 214–22.

Lamb, W. F., T. Wiedmann, J. Pongratz, R. Andrew, M. Crippa, J. G. Olivier, D. Wiedenhofer, G. Mattioli, A. Al Khourdajie, J. House, and S. Pachauri (2021). "A Review of Trends and Drivers of Greenhouse Gas Emissions by Sector from 1990 to 2018." *Environmental Research Letters* 16(073005).

Larsen, J., Z. Patterson, and A. El-Geneidy (2013). "Build It. But Where? The Use of Geographic Information Systems in Identifying Locations for New Cycling Infrastructure." *International Journal of Sustainable Transportation* 7(4), 299–317.

Line, L. (2021). "Investigating the Use of OpenStreetMap to Evaluate Cyclists' Preferences. An Assessment into the use of OpenStreetMap to Understand Cyclists' Perceptions, Preferences, and Experiences on the Cycle Network in Copenhagen and Frederiksberg." Master's Thesis. Copenhagen: University of Copenhagen.

Lovelace, R., A. Goodman, R. Aldred, N. Berkoff, A. Abbas, and J. Woodcock (2017). "The Propensity to Cycle Tool: An Open Source Online System for Sustainable Transport Planning." *Journal of Transport and Land Use* 10(1), 505–28.

Lovelace, R., M. Morgan, J. Talbot, and M. Lucas-Smith (2020). "Methods to Prioritise Pop-Up Active Transport Infrastructure." *Findings* 2020(7).

Lowry, M., and T. H. Loh (2017). "Quantifying Bicycle Network Connectivity." *Preventive Medicine* 95, S134–40.

Mahfouz, H., E. Arcaute, and R. Lovelace (2021). A Road Segment Prioritization Approach for Cycling Infrastructure. ArXiv:2105.03712 [physics.soc-ph].

Mattioli, G., C. Roberts, J. K. Steinberger, and A. Brown (2020). "The Political Economy of Car Dependence: A Systems of Provision Approach." *Energy Research & Social Science* 66, 101486.

Mauttone, A., G. Mercadante, M. Rabaza, and F. Toledo (2017). "Bicycle Network Design: Model and Solution Algorithm." *Transportation Research Procedia* 27, 969–76.

McDaniel, S., M. B. Lowry, and M. Dixon (2014). "Using Origin-Destination Centrality to Estimate Directional Bicycle Volumes." *Transportation Research Record* 2430(1), 12–19.

McDonald, N. (2012). "Children and Cycling." City Cycling. Cambridge, MA, London, UK: MIT Press.

Melson, C. L., J. C. Duthie, and S. D. Boyles (2014). "Influence of Bridge Facility Attributes on Bicycle Travel Behavior." *Transportation Letters* 6(1), 46–54.

Mitra, R., R. A. Ziemba, and P. M. Hess (2017). "Mode Substitution Effect of Urban Cycle Tracks: Case Study of a Downtown Street in Toronto, Canada." *International Journal of Sustainable Transportation* 11(4), 248–56.

Mooney, P., P. Corcoran, and A. C. Winstanley (2010). "Towards Quality Metrics for OpenStreetMap." In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS '10. 514–17. New York, NY, USA: Association for Computing Machinery.

Natera Orozco, L. G., L. Alessandretti, M. Saberi, M. Szell, and F. Battiston (2021). Multimodal urban mobility and multilayer transport networks ArXiv:2111.02152, [physics.soc-ph].

Natera Orozco, L. G., F. Battiston, G. Iñiguez, and M. Szell (2020a). "Data-Driven Strategies for Optimal Bicycle Network Growth." *Royal Society Open Science* 7(12), 201130.
Natera Orozco, L. G., F. Battiston, G. Iñiguez, and M. Szell (2020b). "Extracting the Multimodal Fingerprint of Urban Transportation Networks." *Transport Findings* 13171.
National Association of City Transportation Officials (NACTO) (2019). "Don't Give Up at the Intersection." Designing All Ages and Abilities Bicycle Crossings. https://nacto.org/publication/dont-give-up-at-the-intersection/.
Nohr, E. A., and Z. Liew (2018). "How to Investigate and Adjust for Selection Bias in Cohort Studies." *Acta Obstetricia et Gynecologica Scandinavica* 97(4), 407–16.
Okabe, A. (2012). "Spatial Analysis Along Networks: Statistical and Computational Methods." Statistics in Practice. Hoboken, NJ: Wiley.
Olmos, L. E., M. S. Tadeo, D. Vlachogiannis, F. Alhasoun, X. E. Alegre, C. Ochoa, F. Targa, and M. C. González (2020). "A Data Science Framework for Planning the Growth of Bicycle Infrastructures." *Transportation Research Part C: Emerging Technologies* 115, 102640.
OpenStreetMap Contributors (2021). Bicycle—OpenStreetMap Wiki. [Online] https://wiki.openstreetmap.org/wiki/Bicycle, accessed: 2022-01-06.
Palominos, N., D. A. Smith, and S. Griffiths (2021). "Identifying and Characterising Active Travel Corridors for London in Response to Covid-19 Using Shortest Path and Streetspace Analysis." Mapping COVID-19 in Space and Time. Cham: Springer.
Pereira, R. H. M., T. Schwanen, and D. Banister (2017). "Distributive Justice and Equity in Transportation." *Transport Reviews* 37(2), 170–91.
Perrine, K., A. Khani, and N. Ruiz-Juri (2015). "Map-Matching Algorithm for Applications in Multimodal Transportation Network Modeling." *Transportation Research Record* 2537(1), 62–70.
Porta, S., P. Crucitti, and V. Latora (2006). "The Network Analysis of Urban Streets: A Primal Approach." *Environment and Planning B: Planning and Design* 33(5), 705–25.
Poudel, N., and P. A. Singleton (2021). "Bicycle Safety at Roundabouts: A Systematic Literature Review." *Transport Reviews* 41(5), 1–26.
Priya Uteng, T., and J. Turner (2019). "Addressing the Linkages between Gender and Transport in Low- and Middle-Income Countries." *Sustainability* 11(17), 4555.
Rahbek, Vierø A. (2020). "Connectivity for Cyclists? A Network Analysis of Copenhagen's Bike Lanes." Lund: Lund University, Master's Thesis.
Reggiani, G., T. van Oijen, H. Hamedmoghadam, W. Daamen, H. L. Vu, and S. Hoogendoorn (2021). "Understanding Bikeability: A Methodology to Assess Urban Networks." *Transportation* 1–29.
Resch, B., and M. Szell (2019). "Human-Centric Data Science for Urban Studies." *ISPRS International Journal of Geo-Information* 8(584).
Sakshaug, L., A. Laureshyn, A. Svensson, and C. Hydén (2010). "Cyclists in Roundabouts-Different Design Solutions." *Accident Analysis & Prevention* 42(4), 1338–51.
Schonlau, M., A. van Soest, A. Kapteyn, and M. Couper (2009). "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods & Research* 37(3), 291–318.
Skov-Petersen, H., B. Barkow, T. Lundhede, and J. B. Jacobsen (2018). "How do Cyclists Make Their Way?–A GPS-Based Revealed Preference Study in Copenhagen." *International Journal of Geographical Information Science* 32(7), 1469–84.
Sørensen, M. W. J., M. L. Jensen, and W. Hansen (2020). "Evaluering af fremførte og afkortede cykelstier." *Proceedings from the Annual Transport Conference at Aalborg University* 27(1).
Szell, M., S. Mimar, T. Perlman, G. Ghoshal, and R. Sinatra (2021). Growing Urban Bicycle Networks ArXiv:2107.02185, [physics.soc-ph].
Thomas, B., and M. DeRobertis (2013). "The Safety of Urban Cycle Tracks: A Review of the Literature." *Accident Analysis & Prevention* 52, 219–27.
Trap Danmark (2021). Københavns bydele. [Online] https://trap.lex.dk/Kobenhavns_bydele, accessed: 2022-01-06.
U.S. Department of Transportation, Federal Highway Administration (2017). Proven Safety Countermeasures. Roundabouts. https://safety.fhwa.dot.gov/provencountermeasures/roundabouts/.

van Eldijk, J., J. Gil, N. Kuska, and R. Sisinty Patro (2020). "Missing Links-Quantifying Barrier Effects of Transport Infrastructure on Local Accessibility." *Transportation Research Part D: Transport and Environment* 85, 102410.

Vejdirektoratet (2017). "Prevent Right-Turn Accidents." Road and Traffic Engineering Measures in Signalized Intersections. https://www.vejdirektoratet.dk/api/drupal/sites/default/files/publications/prevent_rightturn_accidents.pdf

Vejdirektoratet (2020). "Vejtekniske løsninger for cyklister." Effekt på sikkerhed og oplevet tryghed. https://www.cyklistforbundet.dk/media/xqcfqkdo/vejdirektoratet_2020_vejtekniske-for-cyklister.pdf.

Vybornova, A. (2021). Identifying and Classifying Gaps in the Bicycle Network of Copenhagen. Master's Thesis, University of Copenhagen.

Wagenbuur, M. (2013). Experimental Bicycle Roundabout in Zwolle. [Online] https://bicycledutch.wordpress.com/2013/08/26/experimental-bicycle-roundabout-in-zwolle/, accessed: 2022-01-06.

Wagenbuur, M. (2014). Junction Design in the Netherlands. [Online] https://bicycledutch.wordpress.com/2014/02/23/junction-design-in-the-netherlands/, accessed: 2022-01-06.

Wang, H., H. De Backer, D. Lauwers, and S. K. J. Chang (2019). "A Spatio-Temporal Mapping to Assess Bicycle Collision Risks on High-Risk Areas (Bridges)–A Case Study from Taipei (Taiwan)." *Journal of Transport Geography* 75, 94–109.

Yamaoka, K., Y. Kumakoshi, and Y. Yoshimura (2021). "Local Betweenness Centrality Analysis of 30 European." *Cities* ArXiv:2103.11437, [physics.soc-ph].

Ye, P., B. Wu, and W. Fan (2016). "Modified Betweenness-Based Measure for Prediction of Traffic Flow on Urban Roads." *Transportation Research Record* 2563(1), 144–50.

Yeboah, G., J. Porto de Albuquerque, R. Troilo, G. Tregonning, S. Perera, S. A. K. S. Ahmed, M. Ajisola, O. Alam, N. Aujla, S. I. Azam, K. Azeem, P. Bakibinga, Y.-F. Chen, N. N. Choudhury, P. J. Diggle, O. Fayehun, P. Gill, F. Griffiths, B. Harris, R. Iqbal, C. Kabaria, A. K. Ziraba, A. Z. Khan, P. Kibe, L. Kisia, C. Kyobutungi, R. J. Lilford, J. J. Madan, N. Mbaya, B. Mberu, S. F. Mohamed, H. Muir, A. Nazish, A. Njeri, O. Odubanjo, A. Omigbodun, M. E. Osuh, E. Owoaje, O. Oyebode, V. Pitidis, O. Rahman, N. Rizvi, J. Sartori, S. Smith, O. J. Taiwo, P. Ulbrich, O. A. Uthman, S. I. Watson, R. Wilson, and R. Yusuf (2021). "Analysis of OpenStreetMap Data Quality at Different Stages of a Participatory Mapping Process: Evidence from Slums in Africa and Asia." *ISPRS International Journal of Geo-Information* 10(4), 265.

Zhang, D., D. J. A. V. Magalhães, and X. C. Wang (2014). "Prioritizing Bicycle Paths in Belo Horizonte City, Brazil: Analysis Based on User Preferences and Willingness Considering Individual Heterogeneity." *Transportation Research Part A: Policy and Practice* 67, 268–78.

Zhu, L., and Y.-C. Chiu (2015). "Transportation Routing Map Abstraction Approach: Algorithm and Numerical Analysis." *Transportation Research Record: Journal of the Transportation Research Board* 2528(1), 78–85.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

Data Mining and Knowledge Discovery (2022) 36:448–476
https://doi.org/10.1007/s10618-021-00812-1

Strengthening ties towards a highly-connected world

Antonis Matakos[1] · **Aristides Gionis**[2]

Received: 20 September 2020 / Accepted: 6 November 2021 / Published online: 4 January 2022

Abstract

Online social networks provide a forum where people make new connections, learn more about the world, get exposed to different points of view, and access information that were previously inaccessible. It is natural to assume that content-delivery algorithms in social networks should not only aim to maximize user engagement but also to offer opportunities for increasing connectivity and enabling social networks to achieve their full potential. Our motivation and aim is to develop methods that foster the creation of new connections, and subsequently, improve the flow of information in the network. To achieve our goal, we propose to leverage the *strong triadic closure* principle, and consider violations to this principle as opportunities for creating more social links. We formalize this idea as an algorithmic problem related to the densest k-subgraph problem. For this new problem, we establish hardness results and propose approximation algorithms. We identify two special cases of the problem that admit a constant-factor approximation. Finally, we experimentally evaluate our proposed algorithm on real-world social networks, and we additionally evaluate some simpler but more scalable algorithms.

Keywords Strong triadic closure · STC · Link recommendations · Densest subgraph discovery

1 Introduction

In the past decade we have witnessed social networks becoming an integral part of society. Social networks like Facebook, Twitter and LinkedIn have grown steadily in recent years, attracting billions of users, and becoming a staple in our everyday

Responsible editor: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautman.

✉ Antonis Matakos
antonis.matakos@aalto.fi

[1] Aalto University, Espoo, Finland

[2] KTH Royal Institute of Technology, Stockholm, Sweden

life. Users of these networks are offered new ways of interacting with each other, while discovering new people and creating friendships; people nowadays tend to have hundreds of online connections (Ugander et al. 2011). In reality, however, since meaningful interactions require time and effort, not all connections in a network correspond to strong friendships; in fact, most connections correspond to acquaintances.

The distinction between close friends and acquaintances is an important dichotomy we need to make when studying the dynamic behavior of friendships in a social network. Understanding these dynamics is key for the study of many fundamental network concepts. The strength of ties plays a critical role in how information flows in the network, how people get acquainted with each other, and how the structure of the network evolves over time.

An attractive principle from sociology, which can help us understand the dynamics of the strength of social connections on social networks is the *strong triadic closure* (STC) principle. In simple terms, STC states that if two people in a social network have a close friend in common, then there is an increased likelihood that they will become acquainted at some point in the future (Rapoport 1953). More formally, given a classification of social ties into *strong* and *weak*, STC, in its most rigid form, states that if an individual A has strong ties to individuals B and C, then B and C need to have a tie (either strong or weak) between themselves. Strong triadic closure is an intuitive notion having grounds in sociology (Catton 1962). Furthermore, the experiments of Granovetter (1973) and later Easley and Kleinberg (2010) provided empirical evidence for the validity of STC in real-world social networks.

Recent work on using STC for social-network analysis has mainly focused on inferring the strength of social ties. In one of the first works, Sintos and Tsaparas (2014) search for an assignment of tie strengths, which maximizes the number of strong edges, while ensuring that the STC property is respected over the whole network. Subsequent works refined this methodology by studying less rigid versions of the STC property (Adriaens et al. 2020), as well as considering the interplay with community structure (Rozenshtein et al. 2017).

While these works have initiated the study of STC around algorithmic problems, they use the STC property to infer the strength of ties in a *snapshot* of the network. From this perspective, our work is a departure from the previous ways of thinking about STC. Instead of using STC to characterize a static network, we assume that STC describes a *mechanism* by which new connections are formed. Therefore, by assuming that we already know the tie strength, we propose to leverage this mechanism by making content recommendations that will strengthen some ties and, according to STC, lead to the formation of new ties. The goal is to select the connections that, according to the STC property, increase the potential of new social connections.

Our guiding principles are the following. First, fostering new network connections ensures people have more opportunities to meet and create new friendships, thus, maximizing user engagement. In addition, higher connectivity improves the flow of information in the network. Second, we want to achieve our objective with as little external intervention as possible. By using STC we can organically create new links, by only reinforcing existing links. Finally, as we will demonstrate more clearly further, an edge-strengthening recommendation could have a higher impact on the objective, since a single strengthening could result in the formation of many new edges.

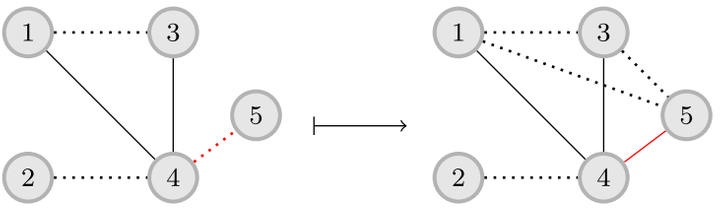


Fig. 1 Illustration of the effect of the STC principle. Solid edges correspond to strong ties and dashed edges to weak ties. Observe (in red) the effect of strengthening tie $(4, 5)$: there is increased chance for ties $(1, 5)$ and $(3, 5)$ to be created. On the other hand, the STC principle does not stipulate creation of tie $(2, 5)$ as $(2, 4)$ is weak (Color figure online)

Our problem formulation is centered around the assumption that according to the STC principle, two people with a common close friend have a higher opportunity to meet and form a new connection. We refer to these connections as STC *bridges*. We also assume that the social network may present opportunities for two people to get to know each other better and strengthen their friendship. Putting together the above ideas, we aim to maximize the number of STC bridges in the network by turning some edges from weak to strong. Note that a single edge might be part of multiple STC bridges, maximizing the potential of strengthening that edge. The example in Fig. 1 demonstrates the effect of an edge strengthening.

We assume that tie strengthening can be achieved in the form of a feature offered by the social network to users who want to opt into. Such a feature would prioritize content from certain users who form weak ties with the user, with the objective of strengthening the tie. It is worth noting that Facebook has experimented with similar ideas, as reported in the media.[1] One can also leverage existing works for strengthening ties in the context of STC (Gilbert and Karahalios 2009; Torro and Pirkkalainen 2017).

We note that our approach is graph-driven instead of user-driven. In particular, we aim to utilize the structure of the social graph, instead of data on user behavior. We believe this ensures that the privacy of the users is better respected, while requiring a minimal amount of data. Additionally this approach leads to an interesting problem formulation, which in turns allows us to develop novel algorithmic ideas. Resulting from this line of thinking we make some modeling assumptions. First, we assume that the question of how to strengthen a tie is specific to a social network, while we aim at an abstract problem formulation that can form a basis for strengthening ties in several different social networks. Therefore, we ask the question: "Given a content recommendation mechanism that has the capability to bring two acquainted people closer together, which ties should we strengthen, in order to maximize the resulting number of connections?" Additionally, in practice it is not equally easy to strengthen each tie. However, since the only distinction we make under STC is into strong and weak ties, we assume that all weak ties can be converted to strong with equal difficulty. Naturally, the trade-off of such an assumption is that it may lead to some recommendations that are uninteresting to the user, given that our approach does not account for user preferences. Additional limitations of our approach and the impact of our assumptions are discussed in more detail in Sect. 10.

[1] https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

Another consideration is that we want to minimize the disruption of the organic structure of the network. To accomplish this objective we consider a limit on how many ties can be converted from weak to strong, by introducing a budget k.

On a more technical level, our problem formulation presents an interesting mapping to a variant of the densest k-subgraph (DkS) problem, which is at the crux of our algorithmic results. From an empirical perspective, we experimentally evaluate our proposed algorithm, in addition to evaluating some simpler but more scalable algorithms.

In summary, we make the following contributions:

- We leverage the strong triadic closure (STC) property in a novel way, for the task of maximally increasing the connections in a social network. We formulate the task as a formal algorithmic problem, which we call MaximizeSTCBridges.
- We prove that the MaximizeSTCBridges problem is **NP**-hard and give approximability results.
- We study the algorithmic properties of our problem in connection to a novel variant of the DkS problem.
- We identify special cases of the problem for which a constant approximation factor can be guaranteed.
- In the experimental section, we propose strong baselines and compare the performance of our algorithm against these baselines.

The rest of the paper is organized as follows. We first put our work in perspective and discuss related work in Sect. 2. Then we present our problem formulation in Sect. 3, while the problem complexity is studied in Sect. 4. In Sect. 5 we reveal the connection of the problem we formulate in this paper with the densest-subgraph problem, and in Sect. 6 we present our algorithm. In Sect. 7 we study properties of the wedge graph, which is used in our construction, and based on these properties, in Sect. 8 we identify two problem variants that admit constant-factor approximation guarantees. In Sect. 9, we present our experimental evaluation for the proposed methods. Finally, in Sect. 10 we discuss limitations of our approach, while Sect. 11 offers a short conclusion and directions for future work.

2 Related work

This paper focuses on leveraging the *strong triadic closure* (STC) property for a novel algorithmic problem. The concept of strong triadic closure was first introduced by Simmel (1908), but it was made popular by Granovetter in his 1973 paper *"the strength of weak ties"* (Granovetter 1973). More recently, the concept was brought again to the forefront in the book of Easley and Kleinberg *"Networks, Crowds and Markets: reasoning about a highly connected world"* (2010), who posit that strong triadic closure occurs in a social network because there is increased opportunity for vertices with a common neighbor to meet, and therefore, create at least weak ties.

Sintos and Tsaparas (2014) study the problem of labeling the edges of the graph to maximize the number of strong edges, such that the assignment satisfies the STC property. Subsequently, Rozenshtein et al. (2017) consider the problem of the inference

of social tie strength, while also taking community structure into account. A recent work by Adriaens et al. (2020) builds directly on the STC-inference problem posed by Sintos and Tsaparas by extending and relaxing their formulation via introducing new constraints and integer labels. While these works use the STC property in order to characterize the ties currently present in the network, we view the STC property as a *process* that takes place in the network and leads to the creation of new edges.

Our paper also shares similarities with other lines of work that consider the introduction of new edges in a social network to improve specific properties. Parotsidis et al. consider adding edges to increase user centrality (Parotsidis et al. 2016), while other works have focused on improving shortest path distance (Meyerson and Tagiku 2009; Papagelis et al. 2011; Parotsidis et al. 2015), diameter (Demaine and Zadimoghaddam 2010), eccentricity (Perumal et al. 2013), communicability (Arrigo and Benzi 2015), and connectivity (Chan et al. 2014). Since in Sect. 8 we consider a problem variant that aims to strengthen so called "local bridges," our work is also similar to the approach of Garimella et al. (2017), who consider the problem of creating bridges to connect communities with opposing views. To the best of our knowledge, this is the first work to take advantage of the STC property for the task of increasing network connectivity.

Central to our work is the well-studied *densest k-subgraph* (DkS) problem. Given a graph G and a parameter k, the DkS problem asks to find a subgraph of G on k vertices with maximum density. The DkS problem has been shown to be **NP**-hard and it does not admit a PTAS under the assumption that **NP** does not contain sub-exponential time algorithms (Khot 2004). The work of Chen et al. (2010), which focuses on the DkS problem on several classes of intersection graphs, provides some essential results for our paper. In particular, our approach relies on adapting their algorithm for a novel variant of DkS, the k-DENSIFY problem.

Drawing further inspiration from the work of Chen et al. we adopt the notion of σ-quasi elimination orders, which generalize perfect elimination orders for chordal graphs. The notion of a σ-quasi elimination order was first proposed by Akcoglu et al. (2002). Ye and Borodin (2009) investigated further the properties of σ-quasi-elimination orders for various graph classes and initiated the study of their algorithmic aspects. Finally, Chen et al. (2010) propose a $\mathcal{O}(\sigma)$-approximation algorithm for the DkS problem if the graph has a polynomial time computable σ-elimination order. In our work we study the σ-quasi elimination properties of a special type of graph that is of interest, and use the properties to derive constant-factor approximation guarantees, in some special cases.

3 Problem formulation

Let $G = (V, E)$ be an undirected graph that represents a social network. The set of vertices V represents individual users, and the set of edges E represents social connections between the individual users. When refering to a subset of vertices $X \subseteq V$ and all edges between them, we will refer to the *induced* subgraph of G, and denote it as $G[X]$.

We consider a labeling ℓ on the edges of the graph, indicating whether each edge $\{v, w\}$ in E corresponds to a *strong* (S) or *weak* (W) social connection. In particular,

this edge labeling is represented as a function $\ell : E \to \{\mathtt{W}, \mathtt{S}\}$. A pair of *incident* edges $e_1 = \{u, v\} \in E$ and $e_2 = \{u, w\} \in E$ where $\{v, w\} \notin E$ is called a *wedge*. We write $(e_1 \wedge e_2)$ to denote the wedge between edges e_1 and e_2. The set of all the wedges in the graph is denoted by W.

The *strong triadic closure* (STC) property states that if a vertex v has strong ties to vertices u and w, i.e., if $\ell(\{v, u\}) = \mathtt{S}$ and $\ell(\{v, w\}) = \mathtt{S}$, then u and w are more likely to form an edge in E, which can be either a weak or a strong tie (Easley and Kleinberg 2010). The absence of the edge $\{u, w\}$, in the presence of strong ties for $\{v, u\}$ and $\{v, w\}$ is called an STC *violation* (Sintos and Tsaparas 2014).

Definition 1 (STC violation) Given a graph $G = (V, E)$ and a labeling function ℓ from the edges of G to $\{\mathtt{W}, \mathtt{S}\}$, a triple of vertices $v, u, w \in V$ constitutes an STC *violation* if $\ell(\{v, u\}) = \mathtt{S}$, and $\ell(\{v, w\}) = \mathtt{S}$, and $\{u, w\} \notin E$. We will denote as $\mathcal{B}(\ell, G)$ the total number of violations on the graph G induced by the labeling ℓ.

The strong triadic closure suggests a structural property that is likely to be true among triples of vertices, but obviously one should not expect it to always hold. A given graph G with a given labeling ℓ may have a large number of violations. In this paper we consider an STC violation as an event that may lead to the formation of new social connections in the graph: two edges $\{v, u\}$ and $\{v, w\}$ with strong ties suggest the possibility for u and w to get acquainted and form a connection. Thus, we view an STC violation as an opportunity for a spontaneous social connection. For this reason we will say that an STC violation leads to an STC *bridge*.

Our goal is to maximize the number of social connections in the network. Since we assume STC bridges will lead to the formation of new edges, we aim to maximize the number of STC bridges. Notice that the network may already contain STC bridges, which however have not yet materialized into new weak edges.

In order to maximize the number of STC bridges we will be looking to convert some edges from weak to strong. We assume this can be achieved through content recommendations for users who opt-in to a feature provided by the social network. An example of such functionality would be to prioritize content generated by a connection in the user's timeline. Such a functionality could present more opportunities for the users to interact with each other. However, as we mentioned, in this paper we focus on the question of *which* ties to select to strengthen, and we consider the problem of *how* to strengthen them, to be orthogonal to our problem. Therefore, for the sake of concreteness and ease of presentation, we consider a simplified setting where each user opts in to receive content recommendations from other users, and also that it is equally difficult to convert each tie, without considering user preferences. Finally, we consider that heavy interference with the natural structure of the network may harm user experience. To amend this, we consider a limit on how many ties can be converted from weak to strong, by introducing a budget k.

Consider two edge labelings ℓ and ℓ' on the graph G. We say that ℓ' is a k-*strengthening* of ℓ if there is a set $E' \subseteq E$ of k edges (i.e., $|E'| = k$) such that (i) for each $\{u, v\} \in E'$ it holds that $\ell(\{u, v\}) = \mathtt{W}$ and $\ell'(\{u, v\}) = \mathtt{S}$, and (ii) for each $\{u, v\} \in E \setminus E'$ it holds that $\ell(\{u, v\}) = \ell'(\{u, v\})$.

Considering the previous discussion we now formulate the problem that we study in this paper.

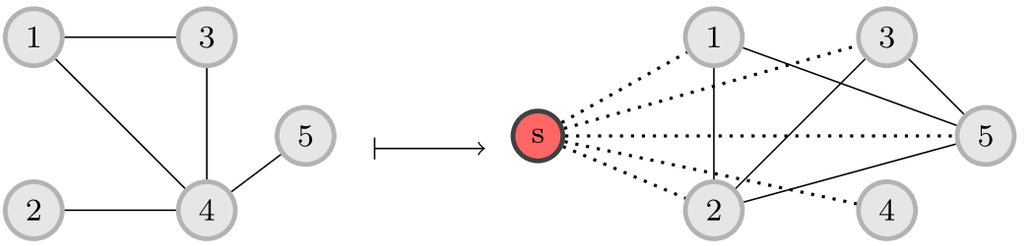


Fig. 2 Construction of graph H used in the proof of Lemma 1

Problem 1 (MAXIMIZESTCBRIDGES) Given a graph $G = (V, E)$ and a labeling function ℓ from the edges of G to $\{\mathtt{W}, \mathtt{S}\}$, find a labeling ℓ' that is a k-strengthening of ℓ and the number of STC bridges $\mathcal{B}(\ell', G)$ induced by ℓ' on G is maximized.

4 Problem complexity

In this section we establish the complexity of Problem 1. We first define the notion of density and formally introduce the densest k-subgraph (DkS) problem.

Definition 2 (Density) Consider an undirected graph $G = (V, E)$. The density of a non-empty subset of vertices $X \subseteq V$ is defined by $\rho(X) = \frac{|E(X)|}{|X|}$, where $E(X)$ is the set of edges in the induced subgraph $G[X]$.

Problem 2 (DkS) Given an undirected graph $G = (V, E)$ and an integer k, the DkS problem asks to find a subset of vertices $S \subseteq V$ such that $|S| = k$ and $\rho(S) \geq c$ (decision version).

We are now ready to show a reduction from the DkS problem to our problem. We will consider a decision variant of MAXIMIZESTCBRIDGES, where we ask for a k-strengthening of the labeling function ℓ' such that $\mathcal{B}(\ell', G) \geq c$. We call this variant MAXIMIZESTCBRIDGES-d. The decision variant can be easily converted into the optimization variant.

Lemma 1 *The problem* MAXIMIZESTCBRIDGES-d *is* **NP**-*complete.*

Proof Given a graph $G = (V, E)$ input to the DkS problem, we create an instance of the MAXIMIZESTCBRIDGES-d problem as follows: we consider the complement of G, which we denote by $\overline{G} = (V, \overline{E})$, and we define by $\{u, v\} \in \overline{E}$ if and only if $\{u, v\} \notin E$. We consider an additional vertex s, which is connected to all other vertices in V. We denote by E_s the set of edges that are incident to s, i.e., $E_s = \{\{s, v\} \mid v \in V\}$. We then construct a new graph $H = (\{s\} \cup V, E_s \cup \overline{E})$. Additionally, we introduce a labeling ℓ so that $\ell(e) = \mathtt{W}$ for all $e \in E_s$ and $\ell(e) = \mathtt{S}$ for all $e \in \overline{E}$. It is straightforward to see that the graph H can be constructed in polynomial time. An example for the construction of graph H can be seen in Fig. 2.

We ask for a solution ℓ' of MAXIMIZESTCBRIDGES-d on the graph H, such that $\mathcal{B}(\ell', H) \geq c$. The first key observation is that no edge $e \in \overline{E}$ can be in the set of strengthened edges in the k-strengthening returned as the solution to the MAXIMIZESTCBRIDGES-d, since it is already $\ell(e) = \mathtt{S}$. Additionally, we observe that

there is no added benefit from any combination of e_1, e_2 with $e_1 \in E_s$, $e_2 \in \overline{E}$, and $\ell'(e_1) = \ell'(e_2) = \mathtt{S}$. This follows from the construction of H, since for any $e_1 = \{s, u\} \in E_s$ and $e_2 = \{u, v\} \in \overline{E}$ it cannot be that $\{v, s\} \notin E_s$. It follows that there cannot exist a wedge $(e_1 \wedge e_2)$ such that $e_1 \in E_s$ and $e_2 \in \overline{E}$.

Let $S \subseteq E_s$ be the edges in E_s that are labeled strong according to ℓ'. Each edge in E_s corresponds to a vertex $v \in V$.

We can see that a pair of selected edges $e_1 = \{s, v\}$, $e_2 = \{s, u\}$ contributes to the objective function of the MAXIMIZESTCBRIDGES-d problem if and only if $\{u, v\} \notin \overline{E}$ and, by construction of H, this happens if and only if $\{u, v\} \in E$. Note also that the MAXIMIZESTCBRIDGES-d problem asks to select k edges in H, which correspond to k vertices in the original graph G. Therefore, the number of STC bridges in H, induced by a solution to the MAXIMIZESTCBRIDGES-d problem is at least c if and only if the corresponding k-subgraph in G has density at least c.

Additionally, given a labeling ℓ' we can verify in polynomial time whether it is a feasible solution for the MAXIMIZESTCBRIDGES-d problem. Therefore, MAXIMIZESTCBRIDGES-d is **NP**-complete. □

Regarding approximability, using the same construction as in the proof of Lemma 1, we can see that a c-approximate solution for MAXIMIZESTCBRIDGES is also a c-approximate solution for the DkS problem. However, the DkS problem has been shown to not admit a PTAS, and the best known approximation ratio to date is $\mathcal{O}(n^{\frac{1}{4}+\epsilon})$ and is due to Bhaskara et al. (2010).

Despite this negative result, in the following sections we show how to obtain a constant-factor approximation guarantee, in polynomial time, for certain special cases of interest.

5 Connection with the densest k-subgraph problem

In the previous section, in order to prove the hardness of the MAXIMIZESTCBRIDGES problem, we reduced the densest k-subgraph (DkS) problem to it. In this section we will delve further into the connection between the two problems, which is a key component of our algorithmic results. Our approach for solving the problem involves the following pipeline: First we transform the input graph into an appropriately constructed *wedge graph*, which maps the problem into a maximum-density finding problem. Then our solution for the MAXIMIZESTCBRIDGES problem is obtained by solving a novel variant of the DkS problem, which we call the *densify k-subgraph* (k-DENSIFY) problem, on the wedge graph.

In Sect. 5.1 we present the k-DENSIFY problem. In Sect. 5.2 we demonstrate how to use the k-DENSIFY problem to solve the MAXIMIZESTCBRIDGES problem, through an appropriately constructed *wedge graph*. In the next section we demonstrate how to solve the k-DENSIFY problem by adapting an existing algorithm for the DkS problem, proposed by Chen et al. (2010), which we briefly describe in Sect. 5.3.

5.1 Densified *k*-subgraph

The k-DENSIFY problem is a variant of DkS, where in addition to the graph G, we also receive as input a set of fixed vertices F, and the goal is to find an additional set S of k vertices that maximize the density of the subgraph induced by the fixed vertices and the newly-selected k vertices. The fixed vertices model the presence of strong edges in the instance of MAXIMIZESTCBRIDGES, which cannot be changed, but still induce STC bridges.

Problem 3 (k-DENSIFY) Given a graph $G = (V, E)$ and a subset of vertices $F \subseteq V$, find a subset of vertices $S \subseteq V$ such that $S \cap F = \emptyset$, $|S| = k$, and the density $\rho(F \cup S)$ of the subgraph induced by the set of vertices $F \cup S$ is maximized.

As one may expect, Problem 3 is **NP**-hard.

Proposition 1 *The k-DENSIFY problem is* **NP***-hard. Furthermore, it does not admit a* PTAS.

Proof It is easy to see that DkS is a special case of k-DENSIFY (by assuming $F = \emptyset$). Furthermore, any approximation algorithm for k-DENSIFY can be used as an approximation algorithm for DkS with the same approximation guarantee. It has been shown that the DkS problem does not admit a PTAS (Bhaskara et al. 2010). □

5.2 The wedge graph

In this section we discuss how to apply k-DENSIFY in order to solve the MAXIMIZESTCBRIDGES problem. Our mapping involves constructing a *wedge graph* $\mathcal{W}$ based on the input graph and solving the k-DENSIFY problem on the wedge graph. The concept of the wedge graph was also used by Sintos and Tsaparas in their work of inferring link types in social networks (Sintos and Tsaparas 2014).

We first present the construction of the wedge graph. Given an input graph $G = (V, E)$ consider the set of wedges W of G, as defined in Sect. 3, i.e., a wedge is a relation between two edges that share a vertex while the "third" edge is not present. The wedge graph of G is a graph $\mathcal{W} = (E, W)$ whose set of vertices are the edges E of G, and whose set of edges are the wedges W of G.

To solve the MAXIMIZESTCBRIDGES problem on the input graph G with a given edge labeling ℓ, we construct the wedge graph $\mathcal{W}$ of G and we take the set of fixed vertices F of $\mathcal{W}$ to be the set of edges of G with strong ties, i.e., $F = \{e \in E \mid \ell(e) = \mathtt{S}\}$. We then solve the k-DENSIFY problem on $\mathcal{W}$ with this set of fixed vertices F, and seeking to find k vertices in $\mathcal{W}$ (i.e., k edges in G). An illustration of this pipeline is shown in Fig. 3.

The solution set S gives a solution for MAXIMIZESTCBRIDGES on G: the selected k vertices maximize the density in S. Since the vertices in $\mathcal{W}$ correspond to edges in G and the edges in the wedge graph correspond to wedges in G, the maximum-density subgraph on the selected vertices in $\mathcal{W}$ corresponds to the set of edges in G that , when relabeled to strong edges in ℓ', maximize the number of STC bridges.

Furthermore, if S is a c-approximate solution for the k-DENSIFY problem, it is also a c-approximate solution for the MAXIMIZESTCBRIDGES problem.

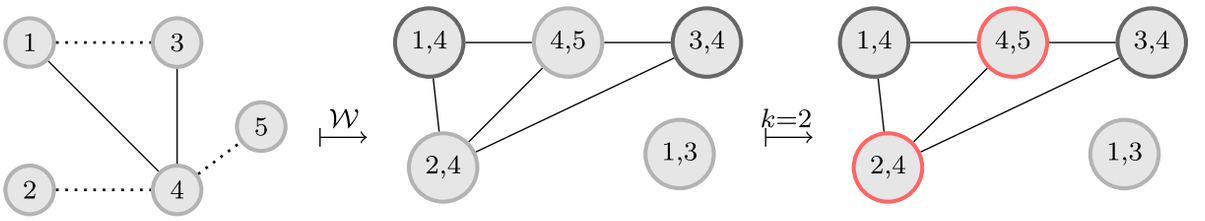


Fig. 3 High-level description of our algorithmic pipeline using the example of Fig. 1 and for $k = 2$. On the left hand side is the initial graph. Then, the graph is transformed into the corresponding wedge graph. Edge-vertices $(1, 4)$ and $(3, 4)$ are highlighted in black since they correspond to fixed vertices in the k-DENSIFY instance (edges $(1, 4)$ and $(3, 4)$ are strong). The final step is to obtain the optimal k-DENSIFY solution for $k = 2$ (vertices in red) (Color figure online)

5.3 Densest k-subgraph algorithm for graphs with σ-quasi-elimination order

A key concept that will be used in our algorithm is the notion of a σ-quasi-elimination order. The concept of σ-quasi-elimination orders was proposed by Akcoglu et al. (2002) as a generalization of perfect elimination orders for chordal graphs. Before formally introducing σ-quasi-elimination orders we introduce some preliminaries.

Let $\alpha(G)$ be the *independence number* of the graph G, i.e., the size of a maximum independent set in G. Let $N(v)$ be the set of neighbors of vertex v, i.e. $N(v) = \{u \mid \{v, u\} \in E\}$. Recall that we denote by $G[S]$ the subgraph of G induced by the vertices of S. If $\mathcal{L} = (v_1, \ldots, v_n)$ is an ordering of the vertices in V, we define $\text{succ}_{\mathcal{L}}(v_i) = \{v_j \mid j > i \text{ and } v_j \in N(v_i)\}$ the set of *successors* of v_i, and $\text{pred}_{\mathcal{L}}(v_i) = \{v_j \mid j < i \text{ and } v_j \in N(v_i)\}$ the set of *predecessors* of v_i. In a perfect elimination order, every set $\text{pred}_{\mathcal{L}}(v_i)$ is a clique. A σ-quasi-elimination order generalizes this definition by relaxing the requirement of having a complete clique.

Definition 3 (σ-quasi-elimination order) Let $G = (V, E)$ be a graph and σ a positive integer. A σ-quasi-elimination order (σ-QEO) of G is an ordering $\mathcal{L}$ of the vertices V such that $\alpha(G[\text{pred}_{\mathcal{L}}(v_i)]) \leq \sigma$, for all $i = 2, \ldots, n$.

We now present the algorithm of Chen et al. (2010) for the DkS problem. The main result of Chen et al. is an $\mathcal{O}(\sigma)$-approximation algorithm for the DkS problem if the input graph has a polynomial-time computable σ-quasi-elimination order.

The algorithm of Chen et al. relies on the *maximum-density subgraph problem* (MDSP) as a key subroutine. The maximum-density subgraph problem is defined as follows: given a graph $G = (V, E, w)$ with non-negative vertex weights $w : V \to \mathbb{R}_{\geq 0}$, we ask to find an induced subgraph $H = (V_H, E_H)$ maximizing the density

$$\rho(H) = \frac{\sum_{v \in V_H} w(v) + |E_H|}{|V_H|}.$$

This problem can be solved optimally in $\mathcal{O}(nm \log(\frac{n^2}{m}))$ time by a reduction to the parametric maximum-flow algorithm (Gallo et al. 1989, Theorem 2.7). The reduction was introduced by Goldberg (1984).

The first step is to solve MDSP on the graph G with weights $w(v) = 0$, for all $v \in V$, and obtain a subgraph H. Let k' be the number of vertices of H. If $k' < k$ then we

repeat the MDSP algorithm on the remaining vertices of G and combine the solution with H. This is Phase 1 of the algorithm, in which we iteratively remove vertices, while keeping track of the number of removed vertices by updating vertex weights in the next call to the MDSP subroutine. If on the other hand $k' > k$, then we are in Phase 2 and some vertices from the obtained solution need to be removed, without losing too much in terms of density.

In the next section, we will adapt the algorithm by Chen et al. for the DkS problem to obtain an algorithm for the k-DENSIFY problem, for which the following proposition holds.

Proposition 2 *Let $G = (V, E)$ be a graph with edge labeling ℓ, and let $\mathcal{W} = (E, W)$ be the wedge graph of G. If $\mathcal{W}$ has a polynomially-time computable σ-QEO, then we can obtain an $\mathcal{O}(\sigma)$-approximation for the* MAXIMIZESTCBRIDGES *problem on graph G. The running time of the algorithm is $\mathcal{O}(\sigma n^{\sigma+2})$.*

6 Proposed algorithm

Our main result in this section is to show that the algorithm of Chen et al. (2010) can be carefully modified in order to solve the k-DENSIFY problem. The resulting algorithm, which we call *sigma-quasi-densify* (SQD), has the same $\mathcal{O}(\sigma)$-approximation guarantee for the k-DENSIFY problem as the algorithm of Chen et al. for the DkS problem.

Now we will describe the main idea behind the modifications of the algorithm, to obtain SQD. Recall that in the k-DENSIFY problem we are given an input graph $G = (V, E)$ and a subset of vertices $F \subseteq V$, and the goal is to find a subset of k vertices $S \subseteq V$ such that the density $\rho(F \cup S)$ of the subgraph induced by the set of vertices $F \cup S$ is maximized.

Let us consider such a solution S to an instance of the k-DENSIFY problem. The density is $\rho = \frac{|E(S)|+|E(F)|+|E(S,F)|}{|S|+|F|}$, where $E(S)$ are the edges contained in the induced subgraph $G[S]$, $E(F)$ are the edges contained in the induced subgraph $G[F]$, and $E(S, F)$ are the edges connecting vertices in S and F. Note that the quantities $|S|$, $|F|$, and $|E(F)|$ are constant, so we can measure the performance of the algorithm with respect to the number of edges in $E(S) \cup E(S, F)$.

To ease the burden of notation, from now on we will write $E' = E(S) \cup E(S, F)$, that is, we disregard the edges among vertices in F when we refer to edges in E'. In addition, we redefine density to be $\rho' = \frac{|E(S)|+|E(S,F)|}{|S|}$. We will refer to the maximum possible value of ρ' as ρ^* and a corresponding set of edges resulting in this density as E^*. We can also see that any c-approximate solution ρ' of ρ^* is also at least c-approximate for k-DENSIFY.

As we will see in more detail below, during Phase 1 of the SQD algorithm we handle the vertices in F by introducing an earlier step before the iterative calls to MDSP, where the fixed vertices are removed and the vertex weights are adjusted accordingly. In Phase 2, we compute a σ-quasi-elimination order for an appropriately constructed graph, which accounts for the vertices in F. By using this newly constructed graph we can prove the following.

Proposition 3 *There exists a $\mathcal{O}(\sigma)$-approximation algorithm for the k-*DENSIFY *problem, if the input graph has a polynomially-time computable σ-QEO.*

We now describe in detail the two phases of our proposed algorithm SQD.
Phase 1: The first phase of the algorithm proceeds by iterative calls to an MDSP subroutine. We introduce a preprocessing step compared to Chen et al. (2010), in which we remove the vertices in F. In particular, we define $G_0 = (V_0, E_0, w_0)$ such that $V_0 = V \setminus F$, $E_0 = E \setminus (E(F) \cup E(V_0, F))$ and $w_0(v) = |E(v, F)|$ for all $v \in V_0$. The rest of Phase 1 proceeds in the same way as in the algorithm of Chen et al. (2010): starting with $i = 0$ we find an optimal solution $H_i = (V_{H_i}, E_{H_i})$ of density ρ_i' by running MDSP on $G_i = (V_i, E_i, w_i)$ with $w_0(v) = |E(v, F)|$ and $w_i(v) = |E(v, U_{i-1} \cup F)|$ in the case $i > 0$, for $v \in V_i$, where $U_i = \cup_{j=0}^{i} V_{H_j}$ is the set of so far all removed vertices (without the vertices in F, and therefore $U_0 = \emptyset$), and $n_i = |U_i|$. Then we form graph G_{i+1} by removing the vertices and incident edges of H_i from G_i. We stop at the first time t such that $n_t \geq \frac{k}{2}$. If $n_t \leq k$, then U_t is returned along with some arbitrary $k - n_t$ vertices Z from V_{t+1} as an approximate solution to the k-DENSIFY problem. Otherwise, if $n_t > k$ we proceed to Phase 2.

We adapt Lemma 4 from Chen et al. (2010), by accounting for the removed vertices from F, to prove that the process described above yields a 4-approximation algorithm.

Lemma 2 (Chen et al. 2010, Lemma 4) *If $n_t \leq k$, the set $U_t \cup Z$ is a 4-approximation solution for the k-*DENSIFY *problem with input graph G.*

The proof of Lemma 2 and all other missing proofs are provided in the Appendix, for better readability.
Phase 2: In this case we have $n_t > k$ and we must delete some vertices from the solution. In order to remove vertices while retaining an approximation guarantee for the quality of the solution, we will use the concept of σ-quasi-elimination orders, which we introduced in the previous section. Recall that a σ-quasi-elimination order (σ-QEO) of G is an ordering $\mathcal{L}$ of the vertices V such that $\alpha(G[\text{pred}_{\mathcal{L}}(v_i)]) \leq \sigma$, for all $i = 2, \ldots, n$. As we will see in Sect. 8.1, computing a σ-quasi-elimination order can be done in time $\mathcal{O}(\sigma^2 n^{\sigma+2})$, using the algorithm presented by Ye and Borodin (2009).

In order to make Phase 2 work similarly to the algorithm of Chen et al. (2010), we need to make a slight modification. Recall that Phase 1 results in a set of vertices U_t. We define $J = G[U_t]$ as the subgraph induced by vertex set U_t. We add some vertices and edges to the induced subgraph J to obtain graph J'. In particular, we construct J' by introducing $|E(v, F)|$ dummy vertices for each vertex $v \in U_t$ and connecting each one of these dummy vertices with v, through an edge. Then, we compute a σ-quasi-elimination order $\mathcal{L}$ for J'. Observe that we can choose $\mathcal{L}$ in such a way that the new dummy vertices come after the vertices in U_t, and hence $\mathcal{L}$ has a prefix that is a σ-QEO for J.

A key observation is that in an optimal MDSP solution H_t with density ρ_t', it holds for every vertex $v \in H_t$ that $w(v) + deg_H(v) \geq \rho_t'$, for otherwise we could delete this vertex to obtain an induced subgraph of higher density. Based on this observation, we have that for any vertex v it is $|\text{succ}_{\mathcal{L}}(v)| \geq \frac{\rho_t'}{2}$ or $|\text{pred}_{\mathcal{L}}(v)| \geq \frac{\rho_t'}{2}$.

Following Chen et al. (2010), we discern two cases. The first case occurs if there exists a vertex $v \in U_t$ with $|\text{pred}_{\mathcal{L}}(v)| \geq \frac{k}{2}$ (Lemma 5). In this case, the predecessor set is large enough to allow us to to efficiently find a subgraph of $\frac{k}{2}$ vertices that is a $\mathcal{O}(\sigma)$-approximation for DkS. If no vertex in $\mathcal{L}$ has a predecessor set of size at least $\frac{k}{2}$, then Lemma 6 of Chen et al. ensures a $\mathcal{O}(\sigma)$-approximation. In Appendix A.1 we show how to obtain bounds analogous to the ones given by Lemmas 5 and 6 of Chen et al. (2010). Note that the proofs need to be modified to work for k-DENSIFY.

Lemma 3 (Chen et al. 2010, Lemma 5) *If there is a vertex $v \in U_t$ with $|pred_{\mathcal{L}}(v)| \geq \frac{k}{2}$, then we can efficiently find a subgraph of $\frac{k}{2}$ vertices in $pred_{\mathcal{L}}(v)$, which is a $\mathcal{O}(\sigma)$-approximation solution for k-*DENSIFY *on G.*

Lemma 4 (Chen et al. 2010, Lemma 6) *If there is no vertex $v \in U_t$ with $|pred_{\mathcal{L}}(v)| \geq \frac{k}{2}$, then we can efficiently find a subset of U_t of size at most k, which is a $\mathcal{O}(\sigma)$-approximation solution for k-*DENSIFY *on G.*

Thanks to our mapping between k-DENSIFY and MAXIMIZESTCBRIDGES, which utilizes the wedge graph $\mathcal{W}$, it is trivial to convert the output of the SQD algorithm in order to obtain a labeling ℓ' that is a solution to MAXIMIZESTCBRIDGES. To construct ℓ' from ℓ, it suffices to set $\ell'(e) = \text{S}$ for all $v_e \in U_t$, which is the set of vertices of the wedge graph $\mathcal{W}$ returned by SQD.

Pseudocode for our method is given in Appendix A.2 as Algorithm 1.

Running Time: The running time of Phase 1 is due to the iterative calls to the MDSP subroutine and is $\mathcal{O}(nm \log(\frac{n^2}{m}))$, while Phase 2 is dominated by the computation of a σ-QEO. The fastest known algorithm to compute a σ-QEO is due to Ye and Borodin (2009) and requires $\mathcal{O}(\sigma n^{\sigma+2})$. We should also note that in the algorithmic pipeline we have described earlier, we run this algorithm on an $\mathcal{O}(m^2)$ wedge graph. Although this asymptotic running time may seem prohibitive for practical applications, these bounds are loose, and as we will see in the following sections, in practice we do not need to run most of the subroutines of the algorithm on really large graphs.

7 Properties of the wedge graph

We now take a closer look at properties of the wedge graph and we derive conditions that lead to better approximation guarantees for our method. We will investigate the properties of the wedge graph with respect to σ-quasi-elimination orders. Before proceeding, we introduce some conventional notation when referring to cliques in graph theory. Let K_t denote a clique of size t, while a *bi-clique* $K_{t,t}$ is a complete bipartite graph $G(U, V, E)$ where $|U| = |V| = t$. We refer to the *clique number* $\omega(G)$ as the maximum t' such that $K_{t'} \subseteq G$. Apart from this standard notation, for convenience we will call two K_t cliques with only one common vertex a *t-bowtie* .

First, we consider upper bounds for σ of an optimal σ-QEO for the wedge graph $\mathcal{W}$. We first present a (naïve) upper bound on $\alpha(\mathcal{W}[N(v_e)])$ for all vertices v_e of $\mathcal{W}$.

Proposition 4 *Let $G = (V, E)$ be a graph and $\mathcal{W} = (E, W)$ its wedge graph. For all vertices $v_e \in E$ of the wedge graph it holds that $\alpha(\mathcal{W}[N(v_e)]) \leq 2(\omega(G) - 1)$.*

Proof Consider an edge $e = \{u, v\}$ of G and the corresponding vertex v_e of $\mathcal{W}$. Denote by E_u the set of edges incident to vertex u, and E_v the set of edges incident to vertex v. Consider first vertex u. Assume that there are $m_u \leq |E_u|$ edges such that $N_G(u) \setminus N_G(v) = m_u$ (they do not form form a triangle with e). These edges form a wedge with e, and therefore the vertices in $\mathcal{W}$ corresponding to these edges will all be connected with an edge to v_e in $\mathcal{W}$. Additionally, the edges in E_u that pairwise form triangles with a third edge not in E_u, will not be connected with an edge in $\mathcal{W}$. We can see that the maximum independent set of vertices in $\mathcal{W}$ corresponding to edges in E_u is formed when the endpoints of the edges outside of u, form a clique. Since the largest clique in G is $\omega(G)$, and since the case for v is symmetric, $\alpha(\mathcal{W}[N(v_e)])$ is at most $2(\omega(G) - 1)$. □

As a consequence we obtain the following lemma.

Lemma 5 *If G is K_t-free, then* SQD *gives a $(2t - 4)$-approximation guarantee.*

Proof If G is K_t-free, then for all vertices $v_e \in E$ of the wedge graph $\mathcal{W}$ we have $\alpha(\mathcal{W}[N(v_e)]) \leq 2t - 4$. Any arbitrary quasi-elimination order $\mathcal{L}$ will result in $\alpha(\mathcal{W}[\text{succ}_{\mathcal{L}}(v_e)]) \leq 2t - 4$, therefore $\sigma \leq 2t - 4$. □

For example, if G is triangle-free, we obtain a 2-approximation, while if K_3 is the largest clique of G then we have a 4-approximation.

We can see that this bound is not very tight, since a large graph may have a high clique number. Additionally, even the presence of a large clique does not necessarily imply a large lower bound for the value of σ for which a σ-quasi-elimination order of $\mathcal{W}$ exists. As an example, consider the wedge graph $\mathcal{W}$ of a graph that has one (or more) cliques K_t, with $t \geq 3$, but without two distinct K_t cliques sharing a common vertex. Then the wedge graph $\mathcal{W}$ has a σ-quasi-elimination order with $\sigma = 2$.

In the following we will describe a particular subgraph that appears in all graphs for which a σ-quasi-elimination order exists.

First, we have the following theorem related to σ-quasi-elimination orders (proof in Appendix A.1). The theorem presents an upper bound on the σ-quasi-elimination order of a subgraph G induced by a set of vertices S, based on the cardinality $|S|$.

Theorem 1 *Let $G = (V, E)$ be a graph that does not have a σ-quasi-elimination order with $\sigma \leq t$, for some value $t \leq n$. Then there exists a connected subgraph of G induced by a minimal set of vertices $S \subseteq V$, such that $\alpha(G[S]) > t$, and additionally $|S| \geq 2t$.*

An example of such a minimal subgraph such that G does not have a t-QEO is the $K_{t,t}$ bi-clique. We can see that in a $K_{t,t}$ bi-clique all vertices v have $\alpha(G[v \cup N(v)]) = t$, therefore there cannot be a possible elimination ordering of the vertices that produces $\alpha(G[u \cup N(u)] < t$ for any $u \in K_{t,t}$.

The previous results indicate that if a graph does not have a σ-quasi elimination order such that $\sigma \leq t$, there must be a $2t$-star present in G with all edges incident to a K_t clique in only one endpoint. However, here we will prove the following claim about the wedge graph $\mathcal{W}$ (proof in Appendix A.1).

Theorem 2 *If $\mathcal{W}$ does not have a $2t$-QEO then G contains two K_t cliques overlapping in only one vertex.*

Theorem 2 can help us bound the value of the σ for which a σ-QEO exists, in cases where the maximum degree of G is also bounded. Namely, we have the following lemma as a consequence.

Lemma 6 *The wedge graph $\mathcal{W}$ of a graph G with maximum degree Δ has a quasi-elimination order with $\sigma \leq \Delta$.*

Proof Since the maximum degree is Δ, each bowtie can consist of at most two $K_{\Delta/2}$ cliques (see proof of Theorem 2), therefore the graph has at most a Δ-QEO. □

Based on the observations in this section, in the following section we introduce two problem variants with constant factor-approximation guarantees.

8 Constant-factor approximation for special cases of interest

In this section we present two special cases of the MaximizeSTCBridges problem. We show that in both cases the proposed approach provides a constant-factor approximation guarantee. Both special cases restrict the family of input graphs or the choices for the output, yet the restricted problem formulations are motivated by realistic scenarios.

8.1 Graphs with bounded maximum degree

The strengthening of a social connection is a process that requires time and effort by both participants. Additionally, we would like to avoid overwhelming users with too many recommendations and potentially harming their experience. In the first special case we make the assumption that for each vertex we consider only the strongest (but still weak) connections to make stronger. Therefore, we consider strengthening a tie only if it belongs in the top-d weak ties of both vertices. Notice here that we assume that we are able to rank all the edges incident to a vertex in order of their strength. The top-d weak ties of a vertex v, ordered by their strength, are denoted by $P_d(v)$. The set of weak edges that belong in the P_d list of both of their endpoints is denoted by $C_d = \{\{u, v\} \in E \mid \ell(\{u, v\}) = \mathtt{W} \text{ and } \{u, v\} \in P_d(u) \cap P_d(v)\}$.

The problem we consider in this case is the following.

Problem 4 (MaximizeSTCBridges$_d$) Given a graph $G = (V, E)$ and a labeling function ℓ from the edges of G to $\{\mathtt{W}, \mathtt{S}\}$, find a labeling ℓ', which is a k-strengthening of ℓ, and the number of STC bridges $\mathcal{B}(\ell', G)$ induced by ℓ' on G is maximized. Furthermore, the edges that are relabeled to $\mathtt{S}$ by ℓ' are restricted to be in the set C_d.

For this restricted case we have the following approximability result.

Theorem 3 SQD *returns a solution that is guaranteed to be a d-approximation to the* MaximizeSTCBridges$_d$ *problem.*

The proof is an immediate consequence of the results in the previous section.

Proof We saw that the wedge graph $\mathcal{W}$ of a graph G with maximum degree d has a QEO with $\sigma \leq d$. It is easy to see that is the case for the restricted MAXIMIZESTCBRIDGES$_d$ problem class. Therefore, SQD yields a factor-d approximation guarantee. □

Additionally, we can also see that for this problem case we can compute a quasi-elimination order more efficiently.

Ye and Borodin (2009) studied the algorithmic properties of quasi-elimination orders. Among other results, they present an $\mathcal{O}(\sigma^2 n^{\sigma+2})$ algorithm for computing a σ-QEO. Here we show that in practice, we can do better than that.

Theorem 4 *Let G be a graph with n vertices, m edges, and maximum degree Δ. A σ-QEO can be computed in time $\mathcal{O}(\sigma^2 n \Delta^{\sigma+1})$.*

Proof Our construction is similar to the one by Ye and Borodin. We build the bipartite graph $G^* = (A, B)$ as follows. We construct a subset-node in A for each subset of size σ in G and a vertex-node in B for each vertex in G. Observe that since the maximum degree in G is Δ, a vertex in G is connected to at most Δ vertices. Therefore for each vertex we need only construct $\Delta^{\sigma+1}$ subset-nodes in A. We connect a vertex-node to a subset-node with a red edge if the vertex in the vertex-node is adjacent to all vertices in the subset-node, and the vertices in the subset are independent. We connect a vertex-node to a subset-node with a black edge if the vertex in the vertex-node is one of the vertices in the subset-node. Constructing such a graph G^* takes $\mathcal{O}(\sigma^2 n \Delta^{\sigma+1})$. The algorithm is the same as described by Ye and Borodin leading to a total complexity of $\mathcal{O}(\sigma^2 n \Delta^{\sigma+1})$. □

Recall that in this case we consider Δ to be small and thus expect the algorithm to run in reasonable time. We further offer a practical speed up, by observing that we do not need to generate subset-nodes that are not incident to red edges. Notice that a subset-node can never be incident to a red edge if it is not an independent set. Additionally, since all subsets of independent sets are also independent, we can use an apriori-style algorithm, where we generate candidate sets of size k from independent sets of size $k-1$. As we will see in the experiments, this leads to an algorithm that is quite efficient in practice.

8.2 Strengthening local bridges

We now propose another special case of the problem with a different aim—we want to leverage the STC property to increase the number of connections between weakly connected parts of the social network. This can be beneficial in settings where the social network is fragmented into strong communities with a small number of connections between them. Connections between different communities has been shown to be facilitated by *local bridges*, which Easley and Kleinberg define as follows (Easley and Kleinberg 2010).

Definition 4 (Local bridge) An edge between two vertices u and v is called *local bridge* if u and v have no common neighbors.

Local bridges are important because they provide their endpoints with access to parts of the network, and hence sources of information, which would otherwise be far away.

Considering the above discussion we focus on strengthening local bridges. We aim to strengthen the local bridges that according to the STC property will lead to the highest amount of new edges, thus increasing the number of connections between parts of the graph that are weakly connected.

Given a graph $G = (V, E)$, let $L \subseteq E$ be the local bridges of G.

Problem 5 (MAXIMIZELOCALBRIDGES) Given a graph $G = (V, E)$ and a labeling function ℓ from the edges of G to $\{\mathtt{W}, \mathtt{S}\}$, find a labeling ℓ', which is a k-strengthening of ℓ, and the number of STC bridges $\mathcal{B}(\ell', G)$ induced by ℓ' on G is maximized. Furthermore, the edges that are relabeled to $\mathtt{S}$ by ℓ' are restricted to be in the set of local bridges L.

Again for this case, we can show that our algorithm provides a constant-factor approximation guarantee.

Theorem 5 *The* MAXIMIZELOCALBRIDGES *problem has a factor-2 approximation guarantee.*

Proof Observe that in the construction of the wedge graph $\mathcal{W}$ we can ignore all edges $e = \{v_1, v_2\}$ for which $\ell(e) = \mathtt{W}$ and $|N(v_1) \cap N(v_2)| > 1$, since they will not be considered in the in the set of edges to be relabeled by MAXIMIZELOCALBRIDGES, and they cannot create any STC bridges. To obtain the solution we run SQD with input the wedge graph $\mathcal{W}$ and $F = \{e : \ell(e) = \mathtt{S}\}$. Consider an edge $e = \{v_1, v_2\}$ corresponding to a vertex v_e of $\mathcal{W}$ returned by SQD. In the wedge graph $\mathcal{W}$, a vertex v_e corresponding to such edge e will be connected to all edge-vertices $v_{e'}$ where $e' = \{u, v_1\}$, $u \in N(v_1)$. If such an edge e' is labeled $\mathtt{W}$, then since it cannot be part of any triangles, in $\mathcal{W}$ it will be connected to all other edge-vertices in $\{\{u, v_1\} : u \in N(v_1)\}$. If such an edge e' is labeled $\mathtt{S}$ it will be included in the set of fixed vertices F, in the instance of k-DENSIFY. However, recall that in SQD, vertices in F do not have an impact on the optimal σ-QEO. The case for $N(v_2)$ is symmetric, and from this we can see that $\alpha(\mathcal{W}[v_e \cup N(v_e)]) \leq 2$. Therefore there always exists a σ-QEO with $\sigma = 2$, giving a factor-2 approximation. □

9 Experimental evaluation

The goal of the experiments is to evaluate the performance of the algorithms for the MAXIMIZESTCBRIDGES problem, both in terms of the number of STC bridges achieved, and the running time. The experiments are conducted on real data, and demonstrate the practical efficiency of the algorithms.

9.1 Heuristics

The algorithm presented in the previous section is of theoretical interest, however, it is not always scalable to large graphs, due to the large worst-case complexity of the σ-QEO-computation step.

In this section, we consider alternative algorithms, which are simpler to implement and/or more efficient.

Next, we present two greedy algorithms for MAXIMIZESTCBRIDGES, which scale linearly to the size of the input graph. Additionally, as we will see in our experimental evaluation, the greedy algorithms yield solutions of extremely high quality, in practice.

Local heuristic The first scalable algorithm is a heuristic (we will simply call it Heuristic). It greedily selects to strengthen the ties that are adjacent to the biggest number of strong ties, resulting in the biggest number of STC bridges after strengthening a weak tie. It is a local heuristic because it only considers the local benefit of strengthening a weak edge, without adapting for the incremental benefit of strengthening multiple weak ties. We expect this heuristic to perform well for small values of k. Regarding the running time, we can find the number of adjacent strong edges in $\mathcal{O}(m)$, while we need $\mathcal{O}(m + k \log m)$ for the top-k computation, which is also the overall asymptotic running time of the algorithm.

Greedy As we discussed, the Heuristic algorithm has the drawback of selecting edges independently, ignoring the additive benefit of strengthening pairs of edges. Our second greedy algorithm (named Greedy) overcomes this drawback by selecting edges iteratively and evaluating the gain in the objective function for each new edge. The Greedy algorithm starts with the input ℓ, and in each step finds an edge $\{u, v\}$, which $\ell(\{u, v\}) = \mathtt{W}$, and converts it to strong, $\ell'(\{u, v\}) = \mathtt{S}$. The edge is selected greedily, such that $\mathcal{B}(\ell', G) - \mathcal{B}(\ell, G)$ is maximized. The algorithm continues strengthening edges while the total number of selected edges does not exceed the budget k. For a solution with at most p relabeled edges, the cost of selecting the best candidate in each iterative step is $\mathcal{O}(mp)$. With an efficient implementation, the total running time of the Greedy is $\mathcal{O}(mp^2)$. In typical scenarios we can assume $p \ll m$, making the algorithm very efficient.

9.2 Datasets

For our experimental evaluation we use real-world datasets, where each edge represents a social relation between two individuals. We only consider weighted networks, where the edge weights correspond to an empirical strength of the connection. We use the edge weight as a proxy for tie strength, and in the following experiments, we arbitrarily pick the 70% percentile of edge strength as the separator between strong and weak ties. We assume that all weak ties can be converted to strong.

We use seven different datasets in our experiments: *LesMis*, *KDD*, *Facebook*, *Twitter*, *Telecoms*, *BitCoinAlpha*, and *Retweets*. The datasets first appeared in Adriaens et al. (2018) and Lahoti et al. (2018) and were kindly shared with us by the authors. The networks convey different types of social trust, and have been used in STC literature before. Table 1 shows some statistics about our datasets. The first two columns of the table contain the number of vertices and edges of each network, the following two the number of strong and weak edges, while the last column is the global clustering coefficient of the networks. If T is the number of triangles in a graph then the clustering coefficient is $C = \frac{T}{T+W}$.

Table 1 Dataset statistics

Dataset	Vertices	Edges	Strong	Weak	C
LesMis	77	254	72	182	0.498
KDD	2 738	11 073	3 159	7 914	0.162
Facebook	3 228	4 585	867	3 718	0.056
Twitter	4 185	5 680	1 694	3 986	0.007
Telecoms	8 665	12 132	3 218	8 914	0.002
BitCoinAlpha	3 775	14 120	2 506	11 614	0.078
Retweets	200 073	4 009 548	251 450	3 758 098	–

Distinction into strong and weak is based on the 70-percentile of ground-truth tie strengh

9.3 Performance evaluation

We now proceed to evaluate the proposed algorithms with respect to the number of STC bridges they achieve. The experiments were performed on a machine with 28 GB of RAM and 8 cores. SQD is the algorithm described in the previous sections, Greedy and Heuristic are the two greedy algorithms. In the case of SQD we also report (in parentheses) the lowest value of σ for an elimination order, found during the execution of the algorithm. As noted before, this represents the approximation guarantee of the algorithm. Figure 4 shows the results obtained by the algorithms on all datasets (except *Retweets* where only Heuristic terminates within reasonable time), where k is reported as a fraction of the total number of edges m.

We observe that Greedy in most cases achieves the best performance, followed by SQD. In general the algorithms perform closely to each other, however SQD heavily outperforms the other two on the *BitCoinAlpha* dataset. We believe this may be due to the presence of a large dense component in the wedge graph of this dataset, which SQD is able to detect. Heuristic achieves a good performance for smaller values of k, due to picking first the edges that are adjacent to many strong edges, resulting in immediate benefit. For larger values of k, the effect of these edges vanishes. This is mostly evident in the fact that Heuristic is always the worst performing algorithm for $k = 0.07m$ and larger. We remind that despite the fact that in the tested datasets SQD sometimes fails to achieve a better performance than Greedy, it has a performance guarantee, based on the optimal value of σ. Finally, the results also confirm our expectation of low optimal σ values, in practice.

9.4 Scalability

We also perform a scalability analysis of the algorithms, with the results shown in Fig. 5. SQD was not able to terminate within reasonable time on *Telecoms* for any value of k. We can see that Greedy scales linearly with k, while Heuristic is very scalable since its complexity is logarithmic with respect to k. We note that Heuristic is capable of terminating fast even on *Retweets*, which has more than 4 M edges. On this dataset Heuristic terminated within 0.033 seconds and achieved 1 814 583 potential new edges.

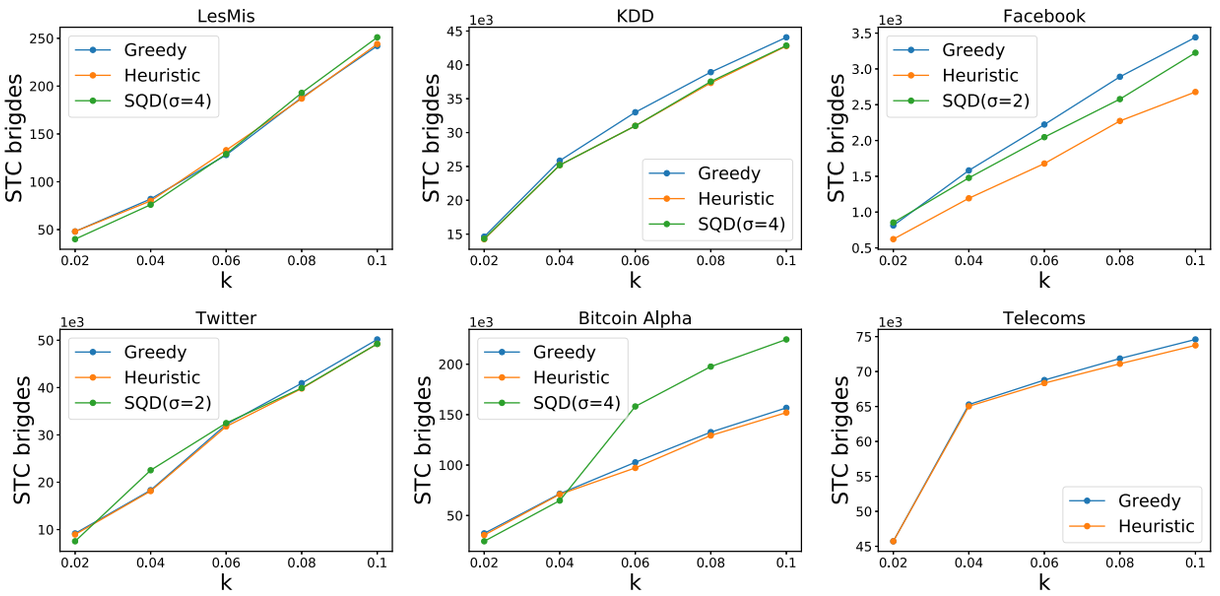


Fig. 4 Performance comparison of all algorithms

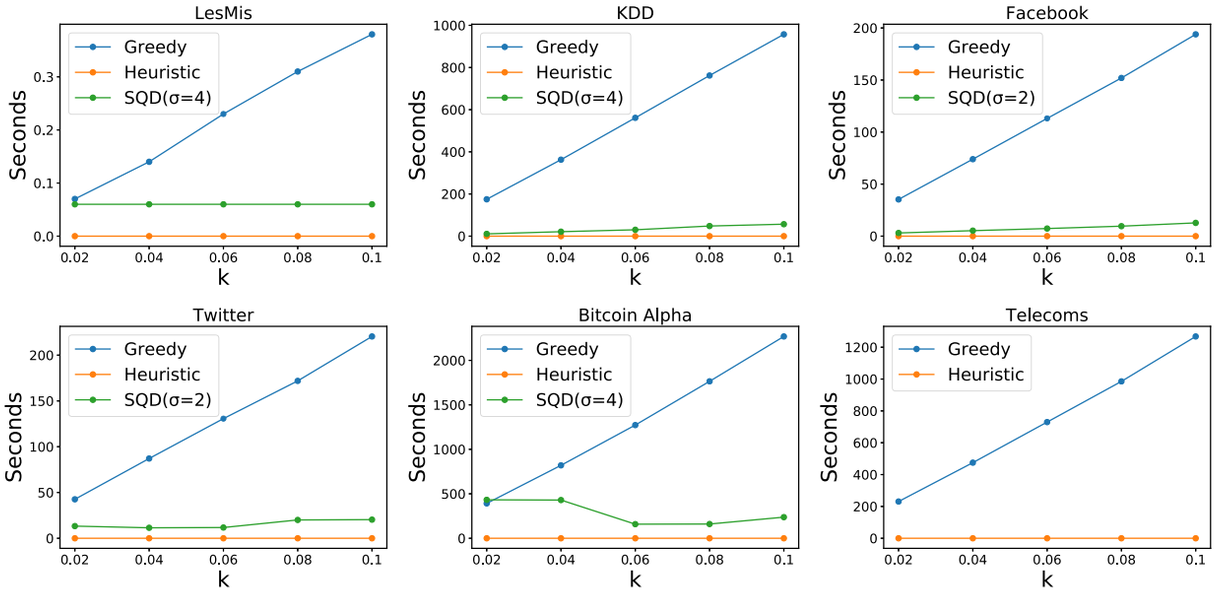


Fig. 5 Running time comparison of all algorithms

Regarding SQD, we notice that it is relatively scalable in most instances and that it does not follow a trend with respect to k. Although the algorithm has a high worst-case complexity, which is dominated by the optimal σ elimination computation step, this step is only applied on the subgraph returned from the first step of the algorithm, which is usually relatively small. We note that the size of the subgraph returned from the first step is dependent on the clustering coefficient of the initial graph; graphs with a low clustering coefficient contain many wedges, and lead to wedge graphs that are denser. For example in *Telecoms*, which has a very low clustering coefficient, the algorithm does not terminate within reasonable time.

10 Limitations and discussion

In this section we discuss limitations of our work. Some of these limitations present challenges to overcome in order to fully realize the potential of our proposed framework, and should serve as a direction for future work.

First, as we noted earlier, our model is graph-driven rather than user-driven. In particular, we aim to utilize the structure of the social graph, while making minimal assumptions regarding user behavior. We only consider relationships between users and distinguish them between strong and weak ties. No other assumption is made about the nature of relationships between users. Accordingly, we have to assume that all weak ties can be converted into strong with equal difficulty. This is an oversimplifying assumption, since in practice not all weak ties are the same, and neither are all strong ties. In order to handle this issue, and given more data on the relationships between users, one could reason about a probabilistic model, where each tie is converted to strong with a certain probability. Note that since in our framework we consider a node-weighted variant of the DkS problem, our solution can be easily adjusted to incorporate the edge-probabilities as node-weights in the wedge graph.

Another implication of our approach to distinguish all edges as only strong or weak, is that it prevents us from giving a specific description of a tie-strengthening mechanism, as such a mechanism would require additional knowledge about the specific nature of friendships in the social network. However, we can assume that such a mechanism is available in the form of a feature that the social network offers to the users to opt-in. We can then only consider strengthening edges whose both endpoints are users that have decided to opt-in to this feature. Note, that we can easily prune from the graph the rest of the edges, and implement our framework on the pruned graph.

It should also be kept in mind that our problem formulation, due to being simple and coarse-grained, can easily be fine-tuned to capture more nuanced cases. Our method describes an algorithm to strengthen edges in a social network instance, but is agnostic of the impact of social connections or how the given graph was created. One may preprocess the graph, for example, and remove all low-strength connections that have low chance of becoming strong (see Sect. 8), so as to focus only on the strongest (but still weak) connections. We believe that our framework offers many capabilities for such fine tuning, which can be incorporated as an additional preprocessing step in the input graph generation.

Another potential limitation of our work is that the proposed algorithm has a step with running time $\mathcal{O}(\sigma^2 n^{\sigma+2})$. Although this may appear infeasible in practice due to the exponentiated σ, in Sect. 7 we show that for high values of σ to be possible, a very specific structure needs to appear in the graph (we call this structure a bowtie). Therefore, even for high-degree and power-law graphs it is unlikely that such large structures will emerge. We empirically demonstrate in the experiments that our algorithm is capable of running even on large datasets. However, developing a faster algorithm is an open line of research that has attracted considerable attention recently.

Finally, in order to minimize the disruption of the organic structure of the network, we consider a limit on how many ties can be converted from weak to strong, by introducing a budget k. However, this may still face the problem of overloading a single user with too many suggestions. In order to handle this, apart from the ideas

mentioned in Sect. 8, one could consider an alternative formulation where a per-user budget b is considered. We note that such an alternative formulation is at least as hard as MAXIMIZESTCBRIDGES. To see this, observe that in the reduction used in Lemma 1, the ego-network of a singe node s is sufficient to reduce to problem to DkS. In this case, the per-user budget b can be set to the global budget k.

11 Conclusion

We considered the problem of leveraging the STC property to introduce new edges in a social network. We formally defined the MAXIMIZESTCBRIDGES problem, and we gave **NP**-hardness and approximability results. We defined a novel variant of the well-studied DkS problem, the k-DENSIFY, which we map to MAXIMIZESTC-BRIDGES. This mapping leads to an approximation algorithm, and additionally allows us to prove various properties of the problem. Utilizing this insight, we define two problem variants that have a constant-factor approximation guarantee. Finally, in the experimental section we experiment with our algorithm in practice and we offer some scalable algorithms, which we evaluate on real data.

Our work opens several interesting directions for future work. A main challenge is to devise algorithms that are both scalable and have provable guarantee for the quality of the solution. Speeding up the algorithm of Ye and Borodin for computing quasi-elimination orders is another challenge towards the same end. Another direction is to explore different constraints regarding the problem of strengthening ties. The present formulation tends to bias a solution towards high degree nodes, which may be undesirable for the behavior of a content-recommendation algorithm. To counter-act this, one may impose a per-user budget for content recommendations.

Finally, another direction for future work is to deploy the proposed algorithm on a real-world social network and evaluate its performance on a practical setting.

Acknowledgements This research is supported by the Academy of Finland Projects AIDA (317085) and MLDB (325117), the ERC Advanced Grant REBOUND (834862), the EC H2020 RIA Project SoBigData++ (871042), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Funding Open Access funding provided by Aalto University.

A Appendix

A.1 Analysis of algorithms

For the analysis of our algorithms we use the following Lemma directly derived from the Turan bound (Chen et al. 2010). Assume that $n = |V|$ and $m = |E|$:

Lemma 7 (Turan bound) *For any graph G,* $m \geq \frac{n^2 - n\alpha(G)}{2\alpha(G)}$

Proof (of Lemma 2) Let $G^*(V^*, E^*)$ be an arbitrary optimal solution to the k-DENSIFY problem. In order to prove the lemma we need to reason about the edges of $G[(U_t \cap V^*) \cup F]$ minus the edges of $G[F]$, since they are not factored in the optimal solution. There are two possible cases. If $|E((U_t \cap V^*) \cup F)| - |E(F)| \geq \frac{|E^*|}{2}$, then $U_t \cup Z$ is trivially a 2-approximation for k-DENSIFY (and hence also a 4-approximation, as the lemma requires). If not, then we define the sets $I_i = U_i \cap V^*$ and $R_i = V^* \setminus I_i$, for all i. We have that $|E((U_t \cap V^*) \cup F)| - |E(F)| = |E(I_t \cup F)| - |E(F)| < \frac{|E^*|}{2}$. Since $|E(I_t \cup F)| = |E(I_t)| + |E(I_t, F)| + |E(F)|$, we have that

$$|E(I_t \cup F)| - |E(F)| < \frac{|E^*|}{2},$$

is equivalent to

$$|E(I_t)| + |E(I_t, F)| + |E(F)| - |E(F)| < \frac{|E^*|}{2},$$

and thus,

$$|E(I_t)| + |E(I_t, F)| < \frac{|E^*|}{2}.$$

Additionally,

$$\begin{aligned}
\rho_i' &= \frac{|E_{H_i}| + |E(U_{i-1} \cup F, V_{H_i})|}{|V_{H_i}|} \\
&\geq \frac{|E(R_{i-1})| + |E(U_{i-1} \cup F, R_{i-1})|}{|R_{i-1}|} \\
&\geq \frac{|E(R_{i-1})| + |E(I_{i-1} \cup F, R_{i-1})|}{|R_{i-1}|} \\
&\geq \frac{|E(R_{i-1})| + |E(I_{i-1}, R_{i-1})| + |E(F, R_{i-1})|}{k}.
\end{aligned}$$

Here we observe that all edges in $|E(F, R_{i-1})|$ belong to E^*, so based on the above we can write:

$$\rho_i' \geq \frac{|E^*| - |E(I_{i-1})| - |E(F, I_{i-1})|}{k}$$

$$\geq \frac{|E^*| - |E(I_t)| - |E(F, I_t)|}{k} \geq \frac{|E^*|}{2k} = \frac{\rho^*}{2}$$

We conclude that

$$|E(U_t)| + |E(F, U_t)| \;\geq\; \min_{i \leq t} \left\{ \rho'_i |U_i| \right\} \;\geq\; \min_{i \leq t} \left\{ \rho'_i \right\} \frac{k}{2} \;\geq\; \frac{|E^*|}{4}.$$

□

Proof (of Lemma 3) We define $\mathcal{A} = \text{pred}_{\mathcal{L}}(v)$. From the σ-quasi-elimination order property, and using Lemma 7, we conclude that the subgraph $G[\mathcal{A}]$ has at least $\frac{1}{2\sigma}\binom{|\mathcal{A}|}{2}$ edges: We choose uniformly at random $\frac{k}{2}$ vertices from $\mathcal{A}$ to form $\mathcal{B}$. Let e be a given edge in $E(\mathcal{A})$. The probability that e is also an edge in the subgraph induced by the randomly-selected set of vertices $\mathcal{B}$ is

$$p_e = \mathbb{P}[e \in E(\mathcal{B})] = \frac{\frac{k}{2}}{|\mathcal{A}|} \frac{(\frac{k}{2} - 1)}{(|\mathcal{A}| - 1)},$$

where $\frac{\frac{k}{2}}{|\mathcal{A}|}$ is the probability that the first endpoint of e is in $\mathcal{B}$, and $\frac{\frac{k}{2}-1}{|\mathcal{A}|-1}$ is the probability that the second endpoint of e is in $\mathcal{B}$ given that the first endpoint is in $\mathcal{B}$. We define the random variable X_e such that

$$X_e = \begin{cases} 1 & \text{if } e \in E(\mathcal{B}), \\ 0 & \text{othersize.} \end{cases}$$

It follows that

$$|E(\mathcal{B})| = \sum_{e \in E(\mathcal{A})} X_e.$$

The expectation of X_e is

$$\mathbb{E}[X_e] = 1 \cdot p_e + 0 \cdot (1 - p_e) = p_e,$$

and thus,

$$\begin{aligned} \mathbb{E}[|E(\mathcal{B})|] &= \mathbb{E}\left[\sum_{e \in E(\mathcal{A})} X_e \right] \\ &= \sum_{e \in E(\mathcal{A})} \mathbb{E}[X_e] \\ &= \sum_{e \in E(\mathcal{A})} p_e \end{aligned}$$

$$= \sum_{e \in E(\mathcal{A})} \frac{\frac{k}{2}}{|\mathcal{A}|} \frac{(\frac{k}{2}-1)}{(|\mathcal{A}|-1)}$$
$$= |E(\mathcal{A})| \frac{\frac{k}{2}}{|\mathcal{A}|} \frac{(\frac{k}{2}-1)}{(|\mathcal{A}|-1)},$$

where the second equality is obtained by the linearity of expectation.

We now apply Lemma 7, which lower bounds $|E(\mathcal{A})| \geq \frac{1}{2\sigma}\binom{|\mathcal{A}|}{2}$ and gives us:

$$\mathbb{E}[|E(\mathcal{B})|] \geq \frac{1}{2\sigma} \frac{|\mathcal{A}|(|\mathcal{A}|-1)}{2} \frac{\frac{k}{2}}{|\mathcal{A}|} \frac{(\frac{k}{2}-1)}{(|\mathcal{A}|-1)} = \frac{1}{\sigma}\left(\frac{k^2}{16} - \frac{k}{8}\right).$$

Notice that we can de-randomize the algorithm using the technique of conditional probabilities.

Additionally, we rank all vertices $v \in G$ according to $|E(v, F)|$ and pick the top $\frac{k}{2}$ vertices to form $\mathcal{C}$. We add the vertices in $\mathcal{C}$ to $\mathcal{B}$, to obtain $\mathcal{B}'$.

We can lower bound $|E(\mathcal{B}')| + |E(\mathcal{B}', F)|$ as

$$\begin{aligned} |E(\mathcal{B}')| + |E(\mathcal{B}', F)| &\geq \frac{1}{\sigma}\left(\frac{k^2}{16} - \frac{k}{8}\right) + |E(\mathcal{C}, F)| \\ &\geq \frac{1}{\sigma}\left(\frac{|E(V^*)|}{16} - \frac{k}{8}\right) + \frac{|E(V^*, F)|}{2} \\ &\geq \frac{1}{\sigma}\left(\frac{k\rho^*}{16} - \frac{k}{8}\right) \end{aligned}$$

We can see that the density of $\mathcal{B}'$ is $\rho'(\mathcal{B}') = \Theta\left(\frac{1}{\sigma}\rho^*\right)$, which is a $\mathcal{O}(\sigma)$-approximation. □

Proof (of Lemma 4) For each vertex v it holds that either $|\text{pred}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2}$ or $|\text{succ}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2}$. Note that $\text{succ}_{\mathcal{L}}(v)$ contains also the copy vertices from F. Since $\text{pred}_{\mathcal{L}}(v)$ does not contain any vertices from F, we can handle the case that $|\text{pred}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2}$, in the same way as Chen et al. (Lemma 6).

We now present the proof of Chen et al., which also works for our case (due to the fact that $\text{succ}_{\mathcal{L}}(v)$ contains the copy vertices from F). We process the vertices of U_t in the reverse order of $\mathcal{L}$, i.e., beginning at v_{n_t}. If a vertex v satisfies the condition $|\text{succ}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2}$, we take just the vertex v, otherwise if it satisfies the condition $|\text{pred}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2}$ we also take a certain subgraph of high-degree vertices of its predecessor set, along with it. To obtain this subgraph we set $\mathcal{A} = \text{pred}_{\mathcal{L}}(v)$ as the starting graph and we repeatedly delete a vertex of degree less than $\frac{|\text{pred}_{\mathcal{L}}(v)|-1}{4\sigma}$. By

$$|\text{pred}_{\mathcal{L}}(v)| \geq \frac{\rho'_t}{2} \geq \frac{\rho^*}{4} \geq 2\sigma,$$

it follows that the induced subgraph $G[\text{pred}_{\mathcal{L}}(v)]$ contains at least $\frac{1}{2\sigma}\binom{|\text{pred}_{\mathcal{L}}(v)|}{2}$ edges, and thus, we cannot delete all vertices (and their edges) of $\text{pred}_{\mathcal{L}}(v)$. We stop if we have collected at least $\frac{k}{2}$ vertices. In every step, we add either a single vertex v or a subset of its predecessors to the solution. Since no vertex has a predecessor set of size at least $\frac{k}{2}$, we select at most k vertices in total, i.e., we obtain a feasible solution $\mathcal{B}$ for k-DENSIFY.

Each vertex v in $\mathcal{B}$ has degree either at least $\frac{\rho'_t}{2}$, if it was selected by the first condition, or it holds that $\frac{|\text{pred}_{\mathcal{L}}(v)|}{4\sigma} \geq \frac{\rho'_t-2}{8\sigma}$, if it was selected by the second condition. Thus,

$$\begin{aligned}|E(\mathcal{B})| + |E(\mathcal{B}, F)| &= \frac{1}{2}\sum_{v\in\mathcal{B}} |E(v,\mathcal{B})| + |E(v,F)| \\ &\geq \frac{\rho'_t - 2}{8\sigma}k \geq \frac{\rho^* - 4}{8\sigma}k \\ &= \mathcal{O}\left(\frac{1}{\sigma}|E^*|\right).\end{aligned}$$

□

Proof (of Theorem 1) Let $S \subseteq V$. If there exists a vertex v in S such that $\alpha(G[v \cup N(v)]) \leq t$ then S is trivially not minimal. Assume for contradiction that there exists a vertex $v \in S$ with $\alpha(G[v \cup N(v)]) > t$, such that $\alpha(G[S]) \leq t$ in the induced subgraph $G[S]$. Then there exists an ordering $\mathcal{L}$ such that all $u_i \notin S \cap N(v)$ are selected before v, so that $\alpha(G[\{v\} \cup \text{succ}_{\mathcal{L}}(v)]) < t$. Then we select v, and since S is connected, this reduces $\alpha(G[\{u\} \cup \text{succ}_{\mathcal{L}}(u)])$ of at least one vertex u in S. Applying this inductively, we obtain a σ-elimination order of S with $\sigma \leq t$. However, this is a contradiction since we assumed that G does not have a σ-elimination order such that $\sigma \leq t$. Therefore this minimal set S has $\alpha(G[S]) > t$.

Additionally, each vertex $v \in S$ must be connected to at least t vertices in S, since S is minimal and $\alpha(G[S]) > t$. Then there exists a vertex $u \in N(v)$ such that $|N(v) \setminus N(v) \cap N(u)| \geq t - 1$, otherwise we would have $\alpha(G[\{v\} \cup S]) \leq t$. This means that the set S has t vertices that are not among the t neighbors of v, therefore $|S| \geq 2t$. □

Proof (of Theorem 2) Recall that a t-bowtie is a set of two K_t's with a single common vertex. First note that if G does not contain a t-bowtie, then it also does not contain a t'-bowtie, with $t' > t$, since a t-bowtie is a subgraph of the t'-bowtie.

Now assume for contradiction that G does not contain a t-bowtie. In the simple case that the graph G does not contain any K_t clique, then it is easy to see that in $\mathcal{W}$ there cannot exist a vertex v_i with $\alpha(\mathcal{W}[v_i \cup N(v_i)]) \geq t$, since there do not exist t pairwise disconnected vertices.

So now we consider the case that G contains K_t cliques but they do not share any vertices. Then for an edge e in G, it is either adjacent to a K_t clique or the corresponding vertex v_e in $\mathcal{W}$ has $\alpha(v_e) < t$. If it is adjacent to a K_t clique then $t \leq \alpha(\mathcal{W}[v_e \cup N(v_e)]) \leq 2t$, however all adjacent edges $e' \in K_t$ have $\alpha(\mathcal{W}[v'_e \cup N(v'_e)]) \leq 2(t-1)$, since they are not adjacent to any K_t. Therefore in an elimination order $\mathcal{L}$ for $\mathcal{W}$ they

can be selected before e, reducing its sequential order number by t. Therefore the optimal σ-quasi-elimination has $\sigma < 2(t-1)$, which is a contradiction.

Now assume that the graph G contains two K_t cliques, with the edges that take part in the K_t cliques denoted as E_c, and there exists an edge e' that forms triangle with two edges from E_c. Then in the wedge graph $\mathcal{W}$ we have $d(v_e) < 2(t-1)$, for $e \in E_c$, and therefore $\alpha(\mathcal{W}[v_e \cup N(v_e)]) < 2(t-1)$.

In all cases we can see that there cannot exist an optimal σ-quasi-elimination with $\sigma \geq 2t$. □

Algorithm 1: SQD

Input: graph G, labeling ℓ, budget k
Output: solution S

$\mathcal{W}(E, W) \leftarrow$WedgeGraph(G) {The wedge graph $\mathcal{W}$ is defined so that its set of vertices are the edges E of G and its set of edges are the wedges W of G.}
$F \leftarrow \{v_e \in E \mid \ell(v_e) = \mathrm{S}\}$
$E_0 \leftarrow E \setminus F$, $W_0 \leftarrow W \setminus W(F)$, $w_0(v_e) \leftarrow |E(v_e, F)| \forall v_e \in E_0$
$\mathcal{W}_0 \leftarrow (E_0, W_0)$
$U_0 \leftarrow \emptyset$; $i \leftarrow 0$
while $|U_i| < \frac{k}{2}$ **do**
  $H_i \leftarrow \mathrm{MDSP}(\mathcal{W}_i, w_i(E))$
  $w_i(v_e) = |E(v_e, U_{i-1} \cup F)| \quad \forall v_e \in E$
  $\mathcal{W}_{i+1} \leftarrow \mathcal{W}_i \setminus H_i$; $U_{i+1} \leftarrow U_i \cup H_i$
  $i \leftarrow i + 1$
end while
$t \leftarrow i - 1$
if $|U_t| \leq k$ **then**
  $Z \leftarrow arbitrary(\mathcal{W}, k - |U_t|)$
  $S \leftarrow U_t \cup Z$
else
  $J' \leftarrow G[U_t] \cup \{(v_e^1, \ldots, v_e^{|E(v_e,F)|}) \quad \forall v_e \in U_t\}$
  $\mathcal{L} \leftarrow \min_\sigma QEO(J', \sigma)$
  if $\exists v_e \in J$ with $|\mathrm{pred}_{\mathcal{L}}(v_e)| \geq \frac{k}{2}$ **then**
    $R(v_e) \leftarrow \mathrm{rank}(|E(v_e, F)|) \quad \forall v_e \in \mathcal{W}$
    $S \leftarrow \mathrm{random\_subset}(\mathcal{W}[\mathrm{pred}_{\mathcal{L}}(v_e)], \frac{k}{2}) \cup \{v_e \mid R(v_e) \leq \frac{k}{2}\}$
  else
    $Q \leftarrow \mathrm{reverse}(\mathcal{L})$; $S \leftarrow \emptyset$
    while $|S| \leq \frac{k}{2}$ **do**
      $S \leftarrow S \cup Q.head$
      if $|\mathrm{pred}_{\mathcal{L}}(Q.head)| \geq \frac{\rho}{2}$ **then**
        $S \cup \mathcal{W}[\mathcal{H}], \mathcal{H} = \{v_e \mid |E(v_e, \mathcal{H})| \geq \frac{|\mathrm{pred}_{\mathcal{L}}(v_e)|-1}{4\sigma}\}$
      end if
      $Q.pop()$
    end while
  end if
end if
return S

A.2 Pseudocode for SQD algorithm

For clarity we provide pseudocode for the SQD algorithm, as Algorithm 1.

References

Adriaens F, Bie TD, Gionis A, Lijffijt J, Rozenshtein P (2018) From acquaintance to best friend forever: robust and fine-grained inference of social tie strengths. CoRR, arXiv:1802.03549

Adriaens F, De Bie T, Gionis A, Lijffijt J, Matakos A, Rozenshtein P (2020) Relaxing the strong triadic closure problem for edge strength inference. Data Min Knowl Discov 34:611–651

Akcoglu K, Aspnes J, DasGupta B, Kao M-Y (2002) Opportunity cost algorithms for combinatorial auctions. In: Kontoghiorghes EJ, Rustem B, Siokos S (eds) Computational methods in decision-making, economics and finance. Springer, Boston, MA, pp 455–479. https://doi.org/10.1007/978-1-4757-3613-7_23

Arrigo F, Benzi M (2015) Edge modification criteria for enhancing the communicability of digraphs. CoRR arXiv:1508.01056

Bhaskara A, Charikar M, Chlamtac E, Feige U,Vijayaraghavan A(2010) Detecting high log-densities: an o(n1/4) approximation for densest k-subgraph. In: STOC, pp 201–210

Catton WR (1962) The acquaintance process. Theodore M. newcomb. Am J Sociol 67(6):704–705

Chan H, Akoglu L, Tong H (2014) Make it or break it: manipulating robustness in large networks. In: SDM. SIAM, pp 325–333

Chen DZ, Fleischer R, Li J (2010) Densest k-subgraph approximation on intersection graphs. In: Jansen K, Solis-Oba R (eds) Approximation and online algorithms. Springer, Berlin, pp 83–93

Demaine ED, Zadimoghaddam M (2010). Minimizing the diameter of a network using shortcut edges. In: SWAT

Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge

Gallo G, Grigoriadis MD, Tarjan RE (1989) A fast parametric maximum flow algorithm and applications. SIAM J Comput 18(1):30–55

Garimella K, Morales GDF, Gionis A, Mathioudakis M (2017) Reducing controversy by connecting opposing views. In: WSDM

Gilbert E, Karahalios K(2009) Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '09, 2009. Association for Computing Machinery, New York, NY, pp 211–220

Goldberg AV (1984) Finding a maximum density subgraph. Technical report, USA

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Khot S (2004) Ruling out PTAS for graph min-bisection, densest subgraph and bipartite clique. In: FOCS, pp 136–145

Lahoti P, Garimella K, Gionis A (2018) Joint non-negative matrix factorization for learning ideological leaning on twitter. In: WSDM

Meyerson A, Tagiku B (2009) Minimizing average shortest path distances via shortcut edge addition. In: APPROX, pp 272–285

Papagelis M, Bonchi F, Gionis A (2011) Suggesting ghost edges for a smaller world. In: CIKM

Parotsidis N, Pitoura N, Tsaparas P (2015) Selecting shortcuts for a smaller world. In: SDM

Parotsidis N, Pitoura E, Tsaparas P (2016) Centrality-aware link recommendations. In: WSDM, pp 503–512

Perumal S, Basu P, Guan Z (2013) Minimizing eccentricity in composite networks via constrained edge additions. In: MILCOM, pp 1894–1899

Rapoport A (1953) Spread of information through a population with socio-structural bias. Bull Math Biophys 15(4):523–533

Rozenshtein P, Tatti N, Gionis A (2017) Inferring the strength of social ties: a community-driven approach. In: KDD, pp 1017–1025

Simmel G (1908) Soziologie Untersuchungen uber die Formen der Vergesellschaftung

Sintos S, Tsaparas P (2014) Using strong triadic closure to characterize ties in social networks. In: KDD, pp 1466–1475. ACM

Torro O, Pirkkalainen H (2017) Strengthening social ties via ICT in the organization. In: Proceedings of the 50th Hawaii international conference on system sciences, pp 5511–5520

Ugander J, Karrer B, Backstrom B, Marlow C (2011) The anatomy of the facebook social graph. CoRR arXiv:1111.4503

Ye Y, Borodin A (2009) Elimination graphs. In: ICALP, pp 774–785

Algorithmic Tools for Understanding the Motif Structure of Networks

Tianyi Chen[1], Brian Matejek[2,3], Michael Mitzenmacher[2], and Charalampos E. Tsourakakis✉[1,2,4]

[1] Boston University, Boston MA, USA
[2] Harvard University, Cambridge MA, USA
[3] Computer Science Laboratory, SRI International, Washington, DC
[4] ISI Foundation, Italy

Abstract. Motifs are small subgraph patterns that play a key role towards understanding the structure and the function of biological and social networks. The current *de facto* approach towards assessing the statistical significance of a motif $\mathcal{M}$ relies on counting its occurrences across the network, and comparing that count to its expected count under some null generative model. This approach can be misleading due to *combinatorial artifacts*. That is, there may be a large count for a motif due to multiple copies sharing many vertices and edges connected to a subgraph, such as a clique, that completes the multiple copies of the motif.

In this work we introduce the novel concept of an (f, q)-spanning motif. A motif $\mathcal{M}$ is (f, q)-spanning if there exists a q-fraction of the nodes that induces an f-fraction of the occurrences of $\mathcal{M}$ in G. Intuitively, when f is close to 1, and q close to 0, most of the occurrences of $\mathcal{M}$ are localized in a small set of nodes, and thus its statistical significance is likely to be due to a combinatorial artifact. We propose efficient heuristics for finding the maximum f for a given q and minimum q for a given f for which a motif is (f, q)-spanning and evaluate them on real-world datasets. Our methods successfully identify combinatorial artifacts that otherwise go undetected using the standard approach for assessing statistical significance.

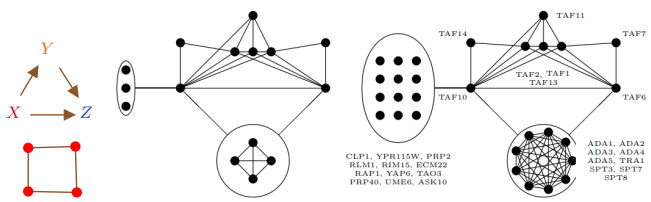
Finally, we leverage the motif structure of a network to design MotifScope, an algorithm that takes as input a graph and two motifs $\mathcal{M}_1, \mathcal{M}_2$, and finds subgraphs of the graph where $\mathcal{M}_1, \mathcal{M}_2$ occur infrequently and frequently respectively. We show that a good selection of $\mathcal{M}_1, \mathcal{M}_2$ allows us to find anomalies in large networks, including bipartite cliques in social graphs, and subgraphs rated with distrust in Bitcoin markets.

Keywords: motifs, graph mining, statistical significance, anomaly detection

1 Introduction

Network motifs, or small induced subgraph patterns, are known to play a key role in understanding the structure and function of various real-world networks,

especially biological [28, 40], and social networks [47]. For example the feed-forward loop (FFL) is one of the most significant subgraphs in the transcription network of the bacteria Escherichia coli. The FFL has three nodes corresponding to transcription factors. The transcription factor X regulates a second transcription factor Y, and together they bind the regulatory region of a target gene Z, jointly modulating its transcription rate [27]. In social networks, triangles (K_3s) are known to appear frequently despite the edge sparsity of the network [49]. Ugander, Backstrom, and Kleinberg [47] showed that on the other hand social networks have very few cycles of length 4 (C_4s). This sheer contrast in the counts of K_3s and C_4s relates to human nature. Specifically, friends of friends are typically friends themselves, thus introducing edges that create K_3s but remove C_4s [49]. An FFL and a C_4 are shown in Figure 1(a).



(a) (b)

Fig. 1: (a) A feed-forward loop (FFL, top) and a C_4 (bottom). (b) Figure source [17]: the subgraph $\mathcal{M}$ on the left appears to be statistically significant in the network G on the right due to the presence of a large independent set, and a large clique in G. The independent set creates $\binom{12}{3}$ stars with three leaves, while the large clique creates $\binom{9}{4}$ smaller cliques of order 4, resulting in a total count of $\binom{12}{3} \times \binom{9}{4}$ occurrences, leading to the misleading conclusion that $\mathcal{M}$ is a statistically significant motif. We refer to this phenomenon as a combinatorial artifact, see also [32, 17].

The *de facto* current approach towards assessing the statistical significance of a motif $\mathcal{M}$ involves two steps: (i) counting the occurrences of $\mathcal{M}$ in the input graph, and (ii) comparing that count to the expected number of occurrences of $\mathcal{M}$ under a null generative model. This approach has been widely used in the literature since the early 2000s [28, 40], but nonetheless has significant drawbacks. The proper choice of the null model is a concern that was raised soon after the publication of the seminal work of Milo et al. [28], see the comment by Artzy et al. [1]. A suitable null model should generate networks similar to the input graph, as otherwise there is a danger of incorrectly assessing a motif as statistically significant (or not) due to an ill-posed null hypothesis. Also importantly, the current approach suffers from *combinatorial artifacts*. As observed originally by Lior Pachter in his blog [32], as well as by Grochow and Kellis [17], the existence of large independent sets and large cliques can obfuscate the relevance of the count of a motif. Consider the motif $\mathcal{M}$ with fifteen nodes corresponding to proteins shown in Figure 1(b) on the left as originally shown in [17]. A node connected with a line to a set of nodes enclosed by a circle/oval denotes that the

node is connected to all the nodes within that set. The closed circle/oval shows the topology of the set of nodes within it. For example, we observe that the node in the middle left is connected to three isolated nodes, whereas the two nodes in the middle (both left and right) are connected to four nodes that form a K_4. Figure 1(b) on the right shows the input network. Due to the existence of a large independent set, and a large clique, the number of occurrences of $\mathcal{M}$ is equal to $\binom{12}{3} \times \binom{9}{4}$. Such a high count may lead to the misleading assessment that $\mathcal{M}$ is statistically significant. Indeed, combinatorial artifacts occur frequently in real-world networks, which often contain large cliques and independent sets, similar to Figure 1(b).

In this work we contribute towards understanding the motif structure of a network (directed or undirected) in the following ways:

• We propose the novel concept of an (f, q)-spanning motif. Specifically, a motif is (f, q)-spanning if there exists a subset of nodes S that induces an f-fraction of the motifs, while being a q-fraction of the node set V. Intuitively, if f is close to 1, and q is close to 0, the motif is likely to be a combinatorial artifact. Based on dense subgraph discovery tools [15], we propose a heuristic algorithm that allows us to test in near-linear time whether a motif is (f, q)-spanning.

• We propose MotifScope, a novel framework that leverages frequently and infrequently appearing motifs to find anomalies in real-world networks. Our framework uses heuristics to find a subgraph that induces many copies of a motif $\mathcal{M}_2$ and few copies of a motif $\mathcal{M}_1$. We show that our framework allows us to find anomalies in social and trust networks.

• We perform an extensive experimental evaluation of various classical and state-of-the-art generative models as null models for assessing statistical significance, which highlights their similarities and differences, as well as the importance of choosing the models.

2 Related Work

Motifs. A motif is typically a subgraph of constant size. The goal of understanding the motif structure of a network spans numerous disciplines, ranging from systems biology [51] to social network analysis [47] and socio-economics [55], as it sheds light into the building blocks of networks [28]. Motifs have found various algorithmic and machine learning applications, under the umbrella of higher order methods [23, 2, 46, 52].

Assessing the statistical significance of a motif. The *de facto* approach for deciding if a motif $\mathcal{M}$ is statistically significant or not relies on comparing its frequency $f_{\mathcal{M}}$ to its expected frequency in a null random graph model [28]. While other approaches to assessing the statistical significance of motifs have been proposed, e.g., [4]; in this work we focus on the prevalent approach as introduced by Milo et al. [28]. Given the null model, one samples a large number of networks with the same number of nodes, and counts the frequency of $\mathcal{M}$; let $\bar{f}_{\mathcal{M}}$, $\sigma_{\mathcal{M}}$ be the average number of occurrences of $\mathcal{M}$ and the sample standard deviation, respectively. The z-score is defined as

$$z\text{-score}(\mathcal{M}) = z_{\mathcal{M}} = \frac{f_{\mathcal{M}} - \bar{f}_{\mathcal{M}}}{\sigma_{\mathcal{M}}}.$$

Observe that the z-score of a motif can be negative; motifs that have a large negative score, and thus appear less often than expected, are sometimes referred in the literature as *anti-motifs* [28, 29].

An important issue is the choice of the null model. A common choice is the configuration model, or one of its variants [5, 14, 10]. This family of models generates a random (di)graph with a given (in-, out-)degree sequence(s). The configuration model was used in the influential works of Milo et al. [28, 29]. However, their approach has received valid critique for a variety of reasons, such as the lack of spatial characteristics [20, 1].

The densest subgraph problem aims to find the subgraph with the maximum average degree over all possible subgraphs [16, 8]. Higher-order extensions have been recently proposed that maximize the average density of a small motif such as a triangle [44, 30]. For this problem, as long as the number of nodes in the small subgraph is constant, there exist both efficient polynomial time exact algorithms [44], and faster greedy approximation algorithms [6, 8].

Graph-based Anomaly Detection is an intensively active area of graph mining [31], with diverse industrial and scientific applications. We discuss related works in greater detail in the Appendix.

3 How to Address Combinatorial Artifacts?

Problem definition. As discussed in Figure 1(b), the significance of the motif on the left hand side does not truly represent statistically significant recurring independent motifs, but rather this motif arises because of a combinatorial artifact [32]. It appears around 30 000 times in a PPI network of S. *cerevisiae*, while its occurrences are concentrated into less than 30 nodes. To help clarify such situations, we provide the following definition.

Definition 1. A motif $\mathcal{M}$ is (f, q)-spanning in graph $G(V, E)$ if there exists a set of nodes $S \subseteq V$ such that $|S| \leq q|V|$ and the induced subgraph $G[S]$ contains an (at least) f-fraction of the occurrences of $\mathcal{M}$ in G.

We will (loosely) say the statistical significance of a motif $\mathcal{M}$ according to some null generative model is a *combinatorial artifact* if it is an (f, q)-spanning motif in $G(V, E)$ with $q \ll 1$, and f close to 1.[5]

[5] It is worth outlining that forcing $f = 1$, and thus simplifying the definition above to a $(1, q)$- or just q-spanning motif is not a robust in the following sense. Consider a graph that is the union of a linear number of node disjoint triangles, and a clique of order $\sqrt{n}$. Each node in the graph participates in a triangle, and thus when $f = 1$, then $q = 1$. However, notice that most of the triangle occurrences appear in the

Our definition of an (f, q)-spanning motif naturally introduces the following optimization problem.

Problem 1. Given a motif $\mathcal{M}$ and a graph $G(V, E)$, what is the largest possible fraction f of occurrences of $\mathcal{M}$ among all subgraphs with (at most) $q|V|$ nodes for a given value of q?

We implicitly assume that the motif $\mathcal{M}$ appears frequently in the graph, and has been assessed statistically significant according to some null generative model; our goal is to understand whether its (apparent) significance is due to a combinatorial artifact or not.

Hardness. Problem 1 is NP-hard, and this holds both when we require $S \subseteq V$ to have exactly $k = q|V|$ nodes, and at most k nodes. The reduction is straightforward, and we omit all details. The idea of the proof is that if we could solve Problem 1, then by setting the motif $\mathcal{M}$ to be a simple undirected edge, we would be able to solve densest-k-subgraph (DkS) problem, and the densest-at-most-k-subgraph (DamkS) problems respectively. Furthermore, we know that these two problems are close in terms of approximation guarantees: if there exists an α-approximation algorithm for the DamkS problem, then there exists an $O(\alpha^2)$ approximation algorithm for the DkS problem. The best known approximation factor for the DkS is $O(n^{-1/4})$ due to Bhaskara et al. [3].

Theorem 1. *Problem 1 is NP-hard.*

We also provide a formulation which aims to optimize q for a given f, stated as the next problem.

Problem 2. Given a motif $\mathcal{M}$ with $m(V)$ total occurrences in a graph $G(V, E)$, what is the smallest possible size $q|V|$ of the union of a set of $f \cdot m(V)$ occurrences for a given value of f?

The results of Chlamtač et al. [9] yield the following corollary.

Corollary 1 (Theorem 1.1 [9]). *Problem 2 is NP-hard. Furthermore, there exists an $O(\sqrt{m(V)})$-approximation algorithm that runs in polynomial time.*

This corollary relates to their results for the minimum p-union problem (MpU). Consider a hypergraph where each hyperedge corresponds to an occurrence of a motif. Problem 2 can be restated as a minimum p-union problem (MpU), with $p = f \cdot m(V)$. However, their approximation algorithm is not practical for our purposes as it relies on computing maximum flows or solving linear

small clique, i.e., $O(\sqrt{n})^3) = O(n^{3/2}) \gg O(n)$. Thus for $f = O(\frac{n^{3/2}}{n+n^{3/2}}) = 1 - o(1)$, q suddenly becomes $O(\frac{\sqrt{n}}{n}) = o(1)$. Similarly, a graph could have multiple distinct smaller combinatorial artifacts, in which case f might be a constant further from 1 (e.g., 3 small subgraphs with each around 1/3 of the motif copies).

programs, and we are interested in motifs with a large number of occurrences. We therefore propose a more efficient heuristic that works for both problem variants.

Algorithm 1: CombArt$(G(V,E), \mathcal{M}, f)$

```
Initialize S*_f = ∅ ;
Count the total number m of occurrences of M in G;
while m(S*_f)/m < f ∧ m(V) > 0 do
    S ← GreedyPeeling(G, M);
    S*_f ← S*_f ∪ S;
    E ← E\E[S*_f] ;
    Update the motif count m(V);
    Compute m(S*_f);
/ * E[S*_f] is the set of edges in the induced subgraph G[S*_f] * / ;
q ← |S*_f|/|V| ;
return q ;
```

Proposed Heuristic. Our heuristic is based on the polynomially time solvable higher-order extension of the densest subgraph problem (DSP) due to Tsourakakis et al. [44, 30]. Our algorithm is shown in pseudocode as Algorithm 1. The algorithm[6] runs as a black-box a greedy peeling algorithm until an f-fraction of the motif occurrences in the graph have been covered by the subgraph $S^\star_f$. In each round, the greedy algorithm provides a $\frac{1}{|V(\mathcal{M})|}$-approximation to the optimization problem $\rho^\star = \max_{S\subseteq V} \frac{m(S)}{|S|}$. Here, $m(S)$ is the number of induced occurrences of motif $\mathcal{M}$ in S. Once the algorithm has covered an f-fraction of $\mathcal{M}$-occurrences in G, we compute q as $|S^\star_f|/n$ where n is the number of nodes in G.

4 MotifScope: Anomaly Detection via Motif Contrasting

A reason statistical significance of motifs is considered a worthwhile issue for study is because it gives us important information about graph structure. Indeed, the existence of subgraphs that occur either frequently or infrequently can have interesting algorithmic implications and applications. Here we consider the problem of using motif counts to determine anomalies in a graph structure, such as a social network. Our results utilize the following natural problem.

Problem 3. Given a frequent motif $\mathcal{M}_1$, and an occurring but infrequent motif $\mathcal{M}_2$ in a graph G, find the subset of nodes $S \subseteq V$ that maximizes the average density difference

$$\max_{S\subseteq V} \frac{m_2(S)}{|S|} - \frac{m_1(S)}{|S|}.$$

[6] While it aims to solve Problem 2, with minor changes it becomes a heuristic for Problem 1.

Intuitively, an induced subgraph $G[S]$ that contains many induced copies of $\mathcal{M}_2$, but few induced copies of $\mathcal{M}_1$ differs significantly from the global network G with respect to those two motifs, and therefore possibly in other interesting ways. To solve Problem 3, we use the dense subgraph discovery framework of Tsourakakis et al. [45] with negative weights. We provide an extension of this approach for contrast of motif structures as follows: each node v is associated with a score $score(v)$ that is equal to $m_2(v) - m_1(v)$. Intuitively, we want to remove nodes that have a large negative score, and keep nodes with a high positive score. The pseudocode is shown in Algorithm 2. Assuming a method MotifCount with time complexity $f(\mathcal{M})$ for motif $\mathcal{M}$, our algorithm runs in $O(n \log n + m + f(\mathcal{M}))$ time in the standard RAM model.

Algorithm 2: MotifScope $(G, \mathcal{M}_1, \mathcal{M}_2)$

$m_i(v) =$# motifs of type $\mathcal{M}_i$ node v is contained in $(i = 1, 2, v \in V(G))$;
$n \leftarrow |V|$;
$H_n \leftarrow G$;
for $i \leftarrow n$ *to* 2 **do**
  Let v be the vertex of G_i of minimum score, i.e.,
    $score(v) = m_2(v) - m_1(v)$ (break ties arbitrarily);
  $H_{i-1} \leftarrow H_i \backslash v$;
  Update counts $m_1(v), m_2(v)$ for all $v \in V$;
return H_j that achieves maximum average density $\frac{m_2(S)-m_1(S)}{|S|}$ among H_is, $i = 1, \ldots, n.$;

Implications and applications. As a specific and important example of the MotifScope algorithm, we explain how it can be used to find dense (near)-bipartite subgraphs. In general, the problem of detecting a dense bipartite subgraph in a graph is NP-hard [25]. Finding such subgraphs is important in practice since large bipartite subgraphs in social and trust networks are known to be rare, and frequently correspond to anomalies, such as a collection of manufactured accounts for illicit uses such as money laundering [43, 33]. To attack this problem using MotifScope we leverage the fact that a bipartite subgraph does not contain any triangles (K_3s), which are otherwise common in social networks, but will probably contain several induced cycles of length 4 (C_4s), which are otherwise rare in social networks [47]. Therefore we set $\mathcal{M}_1 = K_3$ and $\mathcal{M}_2 = C_4$. While our approach is not guaranteed to output a bipartite graph (or even a near-bipartite graph), we show that on real data optimizing for minimizing K_3s while maximizing C_4s often yields a bipartite subgraph in practice. As a rule-of-thumb for using MotifScope for anomaly detection applications, we propose either using prior knowledge of important subgraphs (such as with the K_3 and C_4 example above), or by choosing $\mathcal{M}_1$ to be one of the motifs with high z-score and $\mathcal{M}_2$ to be one of the motifs with low z-score.

5 Experiments

Datasets and Code. Table 1 summarizes the datasets that we use. We use publicly available datasets from a variety of domains, including biological, social, power, and trust networks. The code was written in Python3. We provide both the code and the datasets anonymously at `https://github.com/tsourakakis-lab/motifscope`.

Dataset	$\|V\|$	$\|E\|$	Description	Directed
S. cerevisiae [54]	759	1 593	PPI	×
C. elegans-PPI [54]	2 018	2 930	PPI	×
C. elegans-brain [51]	219	2 416	Connectome	✓
hamsterster [36]	2 426	1 593	Social	×
Eris1176 [36]	1 176	18 552	Power	×
Bitcoin-OTC [22]	5 881	35 592	Trust	✓
Bitcoin-Alpha [21]	3 783	24 186	Trust	✓
LastFM [38]	7 624	27 806	Social	×
Twitch-EN [37]	7 126	35 324	Social	×

Table 1: Summary of datasets.

Experimental Setup. The experiments are performed on a single machine, with an Intel i7-10850H CPU @ 2.70GHz and 32GB of main memory. The motif listing algorithm we use is due to Wernicke [50]. We focus on small-sized subgraphs. Figure 2 presents the 13 possible directed motifs of order 3; we shall refer to each motif with their id, for example $motif_{13}$ is the triangle with all six possible directed edges.

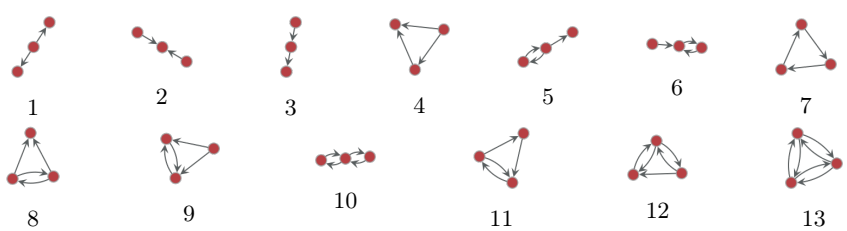


Fig. 2: There exist 13 possible directed motifs of order 3.

5.1 Combinatorial artifacts

Table 2 summarizes the performance of CombArt algorithm on five different networks. The second column of the table visualizes a motif of interest $\mathcal{M}$. We use a similar notation as [17], where a large node annotated as $S - c$ ($K - c$) represents an independent set (clique) with c nodes. We observe that real-world

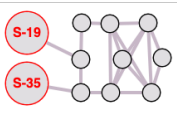
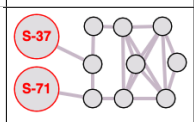
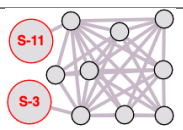
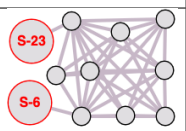
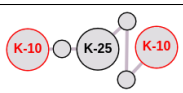
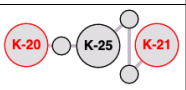
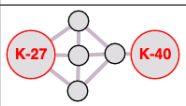
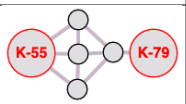
Dataset	Motif	Artifact source	Count	(f, q)
S. cerevisiae			$\binom{37}{19} \times \binom{71}{35}$	$(1, 0.06)$
C. Elegans-PPI			$\binom{23}{11} \times \binom{6}{3}$	$(1, 0.063)$
hamsterster			$\binom{21}{10} \times \binom{20}{10}$	$(1, 0.027)$
Eris1176			$\binom{55}{27} \times \binom{79}{40}$	$(1, 0.117)$
C. Elegans-Brain		-	1554	$(0.8, 0.61)$

Table 2: Motifs that are statistically significant from different networks due to combinatorial artifacts. Subgraphs the motifs are clustered in are also listed together with other statistics.

networks typically contain large cliques and independent sets, and thus there exist various motifs whose significance will be a combinatorial artifact. The third column summarizes the subgraph which causes the combinatorial artifact, while the fourth and fifth columns show the motif count which happens to be also the global count ($f = 1$), and the (f, q) values. As we observe, our novel definition sheds light into assessing the significance of those motifs, by noting that $f = 1$ and q is a small fraction of the node set. In contrast, the FFL motif, which is known to play a biological role, is $(0.8, 0.61)$-spanning, indicating statistical significance is not due to a combinatorial artifact. We believe these examples show our proposed method can be a significant enhancement to the current approach of assessing the statistical significance of motifs.

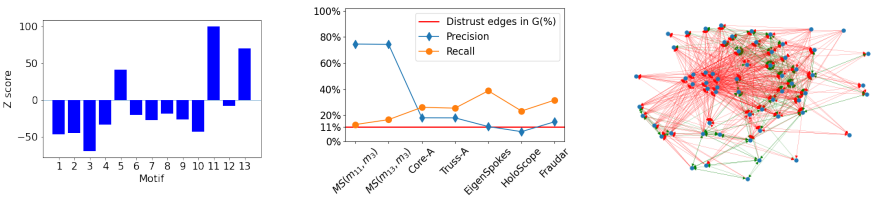
5.2 MotifScope case studies

We show two case studies of MotifScope. The first is an algorithmic application that attacks an NP-hard problem using prior knowledge about the appearance of motifs $\mathcal{M}_1, \mathcal{M}_2$, while the second application first analyzes the network to choose $\mathcal{M}_1, \mathcal{M}_2$.

Bipartite Subgraphs in Social Networks As we mentioned in Section 4, we run MotifScope using $\mathcal{M}_1 = K_3, \mathcal{M}_2 = C_4$, aiming to find a subgraph that induces many cycles of length 4, and few triangles. Our results are summarized in Table 3 for four datasets. We report the total number of induced edges, and

Dataset	# edges	# nodes in L	# nodes in R
LastFM	124	21	37
Bitcoin-Alpha	24	5	9
Bitcoin-OTC	31	6	10
Twitch-EN	61	7	23

Table 3: Bipartite subgraph found by contrasting C_4 and K_3.



(a) (b) (c)

Fig. 3: Results on the Bitcoin-OTC network. (a) When no prior knowledge is available, we use the z-scores. Here, we show the z-scores of the 13 motifs of order 3. (b) Precision and recall for various anomaly detection methods and MotifScope (MS) using as $(\mathcal{M}_1, \mathcal{M}_2)$ motifs $(motif_{11}, motif_3)$, and $(motif_{13}, motif_3)$, see Figure 2 for the actual motifs. (c) Subgraph found by MotifScope for $(motif_{11}, motif_3)$. Distrust relations are colored red, and trust relations are colored green.

the number of nodes in the bi-partition (L, R) of the output node set. Even though our method is not guaranteed to output bipartite subgraphs, the output subgraphs here were in fact all bipartite, i.e., all reported edges having one endpoint in L and one in R.

Anomaly Detection in Trust Networks We use the Bitcoin-OTC network to illustrate the use of MotifScope for anomaly detection on real-world networks. In the Appendix we provide additional results for the Bitcoin-alpha network and camouflage behaviors discovered by MotifScope. Since we have no prior knowledge about the motifs in Bitcoin-OTC, we consider all motifs of order 3, and we compute their z-scores. Figure 3a shows the z-scores of all 13 motifs. We observe that motif 3 has the most negative z-score indicating that it appears significantly less often than what we would expect in the directed configuration model. On the contrary, motifs 11, and 13 appear significantly more often. Thus, we use each of motifs 11 and 13 for $\mathcal{M}_1$, and motif 3 for $\mathcal{M}_2$.

The whole Bitcoin-OTC network contains 11% negative edges, which denote distrust. Figure 3b shows the precision and recall for MotifScope, and popular graph anomaly detection methods that use dense subgraph discovery methods, including Core-A and Truss-A from Corescope [41], EigenSpokes [34], Holoscope [26], and Fraudar [18]. Here, we measure the quality of a subgraph S, using: (i) the precision, namely the fraction of negative edges induced by S over the total number of edges in S, and (ii) the recall, namely the fraction of negative

edges in S over the number of negative edges in the whole graph. We observe that our method outperforms competitors, finding subgraphs that induce a lot of distrust. Figure 3c visualizes one such subgraph. It is worth noting that motifs 11 and 13 are strongly connected, indicating that in this dataset reciprocal edges correlate with trust, whereas motif 3 is a directed chain that lacks reciprocity and correlates with distrust.

Running times. Since our graphs are small to medium size, the main computational bottleneck comes from computing motifs on a large ensemble of sampled graphs from the null models. For instance, for Bitcoin-OTC, listing all motifs of order 3 takes around 20 seconds per sampled graph, and the dense subgraph discovery process (greedy peeling [8]) takes around 17 seconds.

6 Motif Significance and Null Models

As we have seen, the calculation of statistical significance depends on an underlying null model. In this section we study the following questions, to better understand similarities and differences among frequently used null models.

Q1 How robust is the significance (or lack thereof) of a given motif $\mathcal{M}$ across different null models? Is there a consensus between different null models on whether a motif is significant or not?
Q2 What are the sets of motifs that are statistically significant for different null models, and how do these sets compare to each other? How similar are they with respect to ranking motifs according to their z-scores?
Q3 How many samples do we need to generate from a null model, in order to obtain a concentrated estimate of the expected motif count? Is this sample size motif-dependent?

In looking at these questions, We consider seven null models summarized in Table 4 and all 13 motifs of order three in Figure 2. The answer for Q3 is provided in the Appendix due to space constraints. We compare the null models to the well studied *C. elegans* connectome. The network consists of 219 neurons and 2 416 synapses that are represented as nodes and edges respectively, see also Table 1. The network we use corresponds to the adulthood of the *C. elegans*, and was obtained via high-resolution electron microscopy by [51]. All seven generative models we use are well-established in the literature, and they span a period of time from the origins of random graph theory to the most recent advances that involve deep-learning inspired models. Furthermore, we use graph models with independent edge probabilities and dependent edge probabilities. Considering both types of models is important as it was recently shown that random graph models where each edge is added to the graph independently with some probability are inherently limited in their ability to generate graphs with high triangle and other subgraph densities [7]. Furthermore, for any sparse graph, the configuration model is unlikely to generate a large clique. In contrast, it is known that biological networks tend to contain cliques and independent

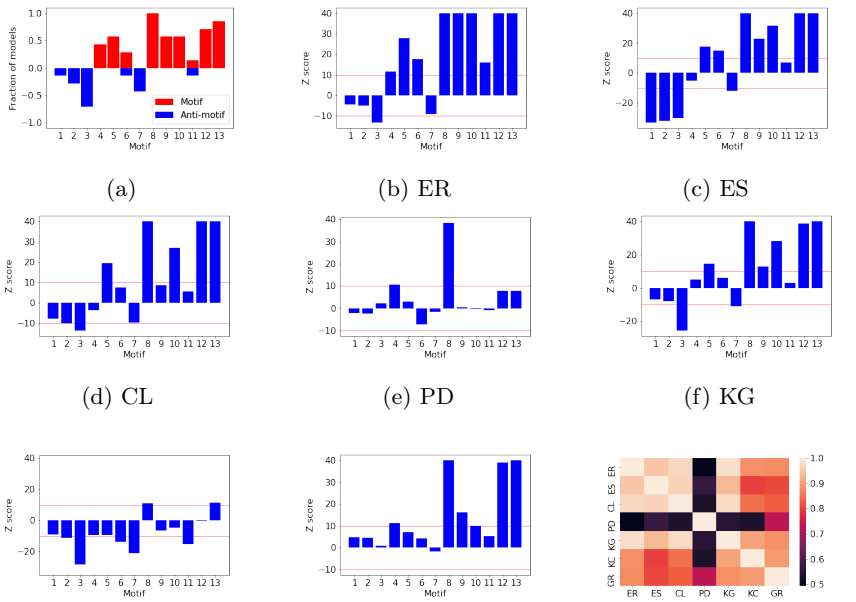
sets [32]. For this reason, we also use state-of-the-art non-independent models including the prescribed k-core model (KC) [48], and GraphRNN [53]. For a detailed description of the models, see the Appendix (supplementary material).

Null Models
Directed Erdős-Rényi model (ER) [13]
Edge swap configuration model (ES) [19]
Chung-Lu model (CL) [11]
Partially directed configuration model (PD) [42]
Stochastic Kronecker graphs (KG) [24]
Prescribed k-core model (KC) [48]
GraphRNN (GRNN) [53]

Table 4: Null models used in our experiments, along with their abbreviation. The first five models are *edge independent*, i.e., each edge $\{i, j\}$ exists independently from the rest with some probability p_{ij}, while KC and GRNN are not.

Is there consensus among null models? Mostly no. We use the *de facto* approach as described in Section 2 to test whether a motif $\mathcal{M}$ appears more often than expected (i.e., $\mathcal{M}$ is a statistically significant motif), or less often than expected (i.e., $\mathcal{M}$ is a statistically significant anti-motif) with respect to each of the seven null models. For each null model, we ensure that we have obtained enough samples for a concentrated estimate of the expectation of each motif $\mathcal{M}$ in Figure 2, by requiring that the coefficient of variation $CV^2 = \frac{\sigma^2_{\mathcal{M}}}{f_{\mathcal{M}}}$ is at most 10^{-2}; the weak law of large numbers guarantees concentration, and is a direct application of Chebyshev's inequality.

For each motif $motif_i, i = 1, \ldots, 13$ we compute the percentage of the null models that assess it as a statistically significant motif (type A), and anti-motif (type B) respectively. Figure 4(a) summarizes our results. For example, motif 11 is assessed as a type A motif by one model, and similarly as type B by one model. According to the five other models, it is not statistically significant in either sense. Figures 4(b)-(g) provide a detailed overview of the assessment of each model. Perhaps surprisingly, motif 8 is the single motif that is assessed as statistically significant by all seven models. Previous research on other *C. elegans* datasets have identified motif 8 as statistically significant in both the male and hermaphrodite sexes [12]. One can construct motif 8 from motif 4, the feedforward loop (FFL), by introducing one reciprocal connection. Analysis of several species has shown that reciprocal connections are over-represented in connectomes [39]. Interestingly, we do not find feedforward loops [28] being statistically significant by several null models, and this can serve as a criterion for the quality of null models but with caution. The absence of several motor neurons in the analyzed connectomes could in part explain the reduced significance of FFLs. There is a general hierarchy of neurons in *C. elegans* with sensory neurons often connecting to interneurons and interneurons often connecting to motor neurons. Although prior research finds the significance of FFLs within each layer, many of the FFLs did contain one neuron of each type [35].



(a) (b) ER (c) ES

(d) CL (e) PD (f) KG

(g) KC (h) GRNN (i) Spearman's correlation

Fig. 4: (a) Histogram of models report each subgraph of size 3 as motif or anti-motif. (b)-(h) Motif significance with respect to z-score by different random graph models. Plots are clipped at a max value of 40. (i) Pairwise Spearman's correlation coefficient of motif z-scores of seven models.

Do null models' rankings agree? Figure 4(i) shows Spearman's correlation coefficient of the z-scores respectively for all pairs of null models. The results are illustrated as a heatmap with the similarity scale on the right. We see that the partially directed configuration model is distinctively different from the rest of the 6 models. We explain this difference due to the fact that *C. elegans* has lots of reciprocal directed arcs, i.e., undirected edges, and thus it can model this aspect better than other models in sparse graphs. We observe that variants of the configuration model are not necessarily similar, a point raised by [14]. GraphRNN produces qualitatively similar results to the partially directed configuration model, but the z-scores are larger due to the fact that the directed version does not capture the frequency of reciprocal edges, despite the wide search of hyperparameters we performed (all details are included in the code).

In a nutshell, caution is required when choosing a null model. Non-independent models, such as the KC and GRNN models, can possibly model complex dependencies that create independent sets and cliques, as described in [7]. GraphRNN seems to be a promising null model for modeling connectomes, although it may not scale well to larger graphs.

7 Conclusion

Understanding the importance of motifs in networks is a key problem in connectomics, with a wide range of applications ranging from social network analysis to machine learning. In this work we introduce the novel concept of an (f, q)-spanning motif that addresses the major issue of *combinatorial artifacts*. We show that determining the smallest value of q for which there exists a node set of cardinality (at most) $q|V|$ that induces an f fraction of the motifs is NP-hard, and we design an efficient heuristic based on dense subgraph discovery methods. Furthermore, we provide new insights into the importance of the null model choice by an extensive empirical analysis of classic and state-of-the-art generative models. Finally, we design the MotifScope framework that uses the motif structure of a graph to detect anomalies.

Our work opens several interesting directions. What are the best non-independent edge models as a null model choice? There is an ongoing line of research, with graph RNNs being a recent example [7, 53]. Can we develop new generative models that leverage motifs for *C. Elegans* and model its temporal evolution, see also [47]?

References

1. Artzy-Randrup, Y., Fleishman, S.J., Ben-Tal, N., Stone, L.: Comment on" network motifs: simple building blocks of complex networks" and" superfamilies of evolved and designed networks". science **305**(5687), 1107–1107 (2004)
2. Benson, A.R., Gleich, D.F., Leskovec, J.: Higher-order organization of complex networks. Science **353**(6295), 163–166 (2016)
3. Bhaskara, A., Charikar, M., Chlamtac, E., Feige, U., Vijayaraghavan, A.: Detecting high log-densities: an $o(n^{-1/4})$ approximation for densest k-subgraph. In: Proc. STOC '10. pp. 201–210 (2010)
4. Bloem, P., de Rooij, S.: Large-scale network motif analysis using compression. Data Mining and Knowledge Discovery **34**(5), 1421–1453 (2020)
5. Bollobás, B.: A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. European Journal of Combinatorics **1**(4), 311–316 (1980)
6. Boob, D., Gao, Y., Peng, R., Sawlani, S., Tsourakakis, C., Wang, D., Wang, J.: Flowless: extracting densest subgraphs without flow computations. In: Proc. TheWebConf '20. pp. 573–583 (2020)
7. Chanpuriya, S., Musco, C., Sotiropoulos, K., Tsourakakis, C.: On the power of edge independent graph models. Advances in NeurIPS **34** (2021)
8. Charikar, M.: Greedy approximation algorithms for finding dense components in a graph. In: APPROX. pp. 84–95. Springer (2000)
9. Chlamt'ač, E., Dinitz, M., Konrad, C., Kortsarz, G., Rabanca, G.: The densest k-subhypergraph problem. arXiv preprint arXiv:1605.04284 (2016)
10. Chung, F., Chung, F.R., Graham, F.C., Lu, L., Chung, K.F., et al.: Complex graphs and networks. No. 107, American Mathematical Soc. (2006)
11. Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. PNAS **99**(25), 15879–15882 (2002)
12. Cook, S.J., et al.: Whole-animal connectomes of both caenorhabditis elegans sexes. Nature **571**(7763), 63–71 (2019)

13. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1), 17–60 (1960)
14. Fosdick, B.K., Larremore, D.B., Nishimura, J., Ugander, J.: Configuring random graph models with fixed degree sequences. Siam Review **60**(2), 315–355 (2018)
15. Gionis, A., Tsourakakis, C.E.: Dense subgraph discovery: Kdd 2015 tutorial. In: Proc. KDD '15. pp. 2313–2314 (2015)
16. Goldberg, A.V.: Finding a maximum density subgraph. University of California Berkeley, CA (1984)
17. Grochow, J.A., Kellis, M.: Network motif discovery using subgraph enumeration and symmetry-breaking. In: Speed, T., Huang, H. (eds.) RECOMB. pp. 92–106 (2007)
18. Hooi, B., Song, H.A., Beutel, A., Shah, N., Shin, K., Faloutsos, C.: Fraudar: Bounding graph fraud in the face of camouflage. In: Proc. KDD '16. p. 895–904 (2016)
19. Kannan, R., Tetali, P., Vempala, S.: Simple markov-chain algorithms for generating bipartite graphs and tournaments. Random Struct. Algorithms **14**(4), 293–308 (1999)
20. King, O.D.: Comment on "subgraphs in random networks". Physical Review E **70**(5), 058101 (2004)
21. Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., Subrahmanian, V.: Rev2: Fraudulent user prediction in rating platforms. In: Proc. WSDM '18. pp. 333–341. ACM (2018)
22. Kumar, S., Spezzano, F., Subrahmanian, V., Faloutsos, C.: Edge weight prediction in weighted signed networks. In: ICDM. pp. 221–230. IEEE (2016)
23. Lee, J.B., Rossi, R.A., Kong, X., Kim, S., Koh, E., Rao, A.: Graph convolutional networks with motif-based attention. In: Proc. CIKM '19. pp. 499–508 (2019)
24. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z.: Kronecker graphs: An approach to modeling networks. J. Mach. Learn. Res (JMLR) **11**, 985–1042 (2010)
25. Lin, B.: The parameterized complexity of the k-biclique problem. Journal of the ACM (JACM) **65**(5), 1–23 (2018)
26. Liu, S., Hooi, B., Faloutsos, C.: Holoscope: Topology-and-spike aware fraud detection. In: Proc. CIKM '17. p. 1539–1548 (2017)
27. Mangan, S., Alon, U.: Structure and function of the feed-forward loop network motif. PNAS **100**(21), 11980–11985 (2003)
28. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science **298**(5594), 824–827 (2002). https://doi.org/10.1126/science.298.5594.824
29. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed networks. Science **303**(5663), 1538–1542 (2004). https://doi.org/10.1126/science.1089167
30. Mitzenmacher, M., Pachocki, J., Peng, R., Tsourakakis, C., Xu, S.C.: Scalable large near-clique detection in large-scale networks via sampling. In: Proc. KDD '15. pp. 815–824. ACM (2015)
31. Noble, C.C., Cook, D.J.: Graph-based anomaly detection. In: Proc. KDD '03. pp. 631–636 (2003)
32. Pachter, L.: Why i read the network nonsense papers. `https://liorpachter.wordpress.com/2014/02/12/why-i-read-the-network-nonsense-papers/`
33. Pandit, S., Chau, D.H., Wang, S., Faloutsos, C.: Netprobe: a fast and scalable system for fraud detection in online auction networks. In: WWW (2007)

34. Prakash, B.A., Sridharan, A., Seshadri, M., Machiraju, S., Faloutsos, C.: Eigenspokes: Surprising patterns and scalable community chipping in large graphs. In: Advances in KDD. pp. 435–448. Springer Berlin Heidelberg (2010)
35. Reigl, M., Alon, U., Chklovskii, D.B.: Search for computational modules in the c. elegans brain. BMC biology **2**(1), 1–12 (2004)
36. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI (2015), `https://networkrepository.com`
37. Rozemberczki, B., Allen, C., Sarkar, R.: Multi-scale attributed node embedding (2019)
38. Rozemberczki, B., Sarkar, R.: Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. In: Proc. CIKM '20. p. 1325–1334 (2020)
39. Scheffer, L.K., et al.: A connectome analysis of the adult drosophila central brain. Elife **9** (2020)
40. Shen-Orr, S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of escherichia coli. Nature genetics **31**, 64–8 (06 2002)
41. Shin, K., Eliassi-Rad, T., Faloutsos, C.: Corescope: graph mining using k-core analysis: patterns, anomalies and algorithms. In: ICDM '16. pp. 469–478 (2016)
42. Spricer, K., Britton, T.: The configuration model for partially directed graphs. Journal of Statistical Physics **161**, 965–985 (2015)
43. Starnini, M., et al.: Smurf-based anti-money laundering in time-evolving transaction networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 171–186. Springer (2021)
44. Tsourakakis, C.: The k-clique densest subgraph problem. In: Proc. WWW '15. pp. 1122–1132 (2015)
45. Tsourakakis, C.E., Chen, T., Kakimura, N., Pachocki, J.: Novel dense subgraph discovery primitives: Risk aversion and exclusion queries. In: Proc. ECML PKDD '19. pp. 378–394. Springer (2019)
46. Tsourakakis, C.E., Pachocki, J., Mitzenmacher, M.: Scalable motif-aware graph clustering. In: Proc. WWW '17. pp. 1451–1460 (2017)
47. Ugander, J., Backstrom, L., Kleinberg, J.: Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In: Proc. WWW '13. pp. 1307–1318 (2013)
48. Van Koevering, K., Benson, A., Kleinberg, J.: Random graphs with prescribed k-core sequences: A new null model for network analysis. In: Proc. TheWebConf '21. p. 367–378 (2021)
49. Wasserman, S., Faust, K., et al.: Social network analysis: Methods and applications (1994)
50. Wernicke, S., Rasche, F.: Fanmod: a tool for fast network motif detection. Bioinformatics **22**(9), 1152–1153 (2006)
51. Witvliet, D.e.a.: Connectomes across development reveal principles of brain maturation. Nature **596**(7871), 257–261 (2021)
52. Yin, H., Benson, A.R., Leskovec, J., Gleich, D.F.: Local higher-order graph clustering. In: Proc. KDD '17. pp. 555–564 (2017)
53. You, J., Ying, R., Ren, X., Hamilton, W.L., Leskovec, J.: Graphrnn: Generating realistic graphs with deep auto-regressive models. In: ICML (2018)
54. Yu, H., et al.: High-quality binary protein interaction map of the yeast interactome network. Science (New York, N.Y.) **322**, 104–10 (09 2008)
55. Zhang, X., Shao, S., Stanley, H., Havlin, S.: Dynamic motifs in socio-economic networks. EPL (Europhysics Letters) **108** (12 2014)

How Inclusive Are Wikipedia's Hyperlinks in Articles Covering Polarizing Topics?

Cristina Menghini
Computer Science
Brown University
Providence, Rhode Island, USA
cristina_menghini@brown.edu

Aris Anagnostopoulos
Sapienza University
Rome, Italy
aris@diag.uniroma1.it

Eli Upfal
Computer Science
Brown University
Providence, Rhode Island, USA
eli_upfal@brown.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671943

***Abstract*—*Wikipedia* relies on an extensive review process to verify that the content of each individual page is unbiased and presents a "neutral point of view." Less attention has been paid to possible biases in the hyperlink structure of Wikipedia, which has a significant influence on the user's exploration process when visiting more than one page. The evaluation of hyperlink bias is challenging because it depends on the global view rather than the text of individual pages.**

In this paper, we focus on the influence of the interconnect topology between articles describing complementary aspects of polarizing topics. We introduce a novel measure of *exposure to diverse information* to quantify users' exposure to different aspects of a topic throughout an entire surfing session, rather than just one click ahead. We apply this measure to six polarizing topics (e.g., *gun control* and *gun right*), and we identify cases in which the network topology significantly limits the exposure of users to diverse information on the topic, encouraging users to remain in a *knowledge bubble*. Our findings demonstrate the importance of evaluating Wikipedia's network structure in addition to the extensive review of individual articles.

***Index Terms*—wikipedia, diversity, web**

I. Introduction

Knowledge on Wikipedia is distributed across articles interconnected via hyperlinks. According to Wikipedia's Linking Manual [1], *"Internal links can add to the cohesion and utility of Wikipedia, allowing readers to deepen their understanding of a topic by conveniently accessing other articles."* Consequently, users are *directly* exposed to an article's content and *indirectly* exposed to the content of the pages it points to.

Wikipedia's pages offer high-quality content with emphasis on an unbiased, neutral point of view (NPOV) [2]–[4], thanks to numerous policies and guidelines [5], [6]. Although it provides tools to support the community for curating pages, it lacks a systematic way to contextualize them within the more general articles' network. Indeed, it is hard to evaluate the extent to which the current hyperlinks satisfy their purpose, especially in connecting articles related to a broad topic.

The majority of users who look for a specific information are likely to find their answers on the first Wikipedia page they are visiting [7], whereas about 20% of Wikipedia's users follow hyperlinks within Wikipedia to develop a broad view of a subject.[1] It is therefore important to investigate whether the link structure leads users to visit pages presenting broad and diverse aspects of their topic of interest. We initiate this important study by concentrating on polarizing topics spanning across multiple articles.

For example, consider the topic *abortion*, which is distributed across multiple articles on Wikipedia. Because of its *polarizing* nature, we recognize pages about events, people, or organizations that are associated either with *pro-choice* or *pro-life*. For instance, the page *Abortion-rights movements* describes organizations related to *pro-choice* view. In this particular page, we identify 15 links pointing to articles about *pro-choice* subjects and only 3 hyperlinks directed to *pro-life* related pages. Furthermore, if we consider articles at distance 2 from the page *Abortion-rights movements*, then there are 4 times more pages associated with *pro-choice* than articles associated with *pro-life* subjects. Similar counting, starting from the page *Anti-abortion movements*, shows 18 outgoing links to *pro-life* pages and only 1 to a *pro-choice* article. At distance two we have 15 times more pages related to the category *pro-life* than pages related to *pro-choice*.

The example above demonstrates unbalanced hyperlink structure in Wikipedia that may influence users' exposure to diverse information on a topic. On another, albeit very different, platform a recent work [9] empirically showed that its recommender system contributes to radicalizing users' pathways. Given the major role of Wikipedia as a popular primary source of knowledge, it is important to evaluate the effect of its hyperlink structure on user navigation, to guarantee a balanced access to well-rounded knowledge.

Evaluating the influence of the hyperlink topology is challenging because it requires a broad view of the network topology, not just the text of a single article. Such a view, and a technique to analyzing it, is not readily available to Wikipedia editors. In this work, we develop an algorithmic approach to quantify users' exposure among a set of articles. Then, we audit the extent to which the current Wikipedia's link structure allows users to browse different stances of polarizing topics.

[1]For the English Wikipedia, the number of unique devices is around 800 million per month. If a device corresponds to a user, around 160 million of users click at least one link throughout their visit on Wikipedia [8].

Our main contributions are the following:

- We initiate the study of the hyperlink network's role in driving users to explore articles of different categories. We investigate this on a set of polarizing topics.
- We design two metrics, the *exposure to diverse information* and the *(mutual) exposure to diverse information*, which quantify the likelihood of visiting pages belonging to different sets of articles click-after-click (Sect. IV).
- We apply our new measures to Wikipedia hyperlink subgraphs related to six polarizing topics. We identify cases in which the Wikipedia's hyperlink network significantly limits the exposure of users to diverse information on the topic (Sect. V).

The code to replicate the analysis is available at https://github.com/CriMenghini/WikiNetBias.

II. Related Works

Improving Wikipedia. Previous works proposed semi-automated procedures to improve Wikipedia's quality by checking the veracity of references [10], [11], suggesting articles' structure [12], looking for hoaxes [13], or recommending links [14], [15]. Among these tools, none provides a measure to evaluate the link-based relationship across articles of diverse categories. In this work, we define such metrics (Sect. IV-B).

Wikipedia Navigation. The literature still lacks a model that generalizes Wikipedia's users' behavior. Previous studies [16]–[19] focused on modeling and predicting human navigation relying on traces from games [19]–[23]. Even though such games provide valuable insights on how users exploit links to move across concepts, other studies showed that users display different behavioral patterns depending on their information needs and the links' position within pages [7], [24], [25]. We exploit such insights to define a general model mimicking localized and in-depth topic exploration (Sect. IV).

Wikipedia Categorization. When dealing with polarizing topics, one needs to distinguish between pages belonging to different side of the topic. Because of the topic granularity, it is hard to rely on automated techniques to categorize articles. Thus, we refrain from using supervised tools as ORES[2] or topic modeling [26], in favor of a mining procedure employed in [27], exploiting actual Wikipedia's categories (Sect. III).

Polarization on Social Media. Many works aim to quantify polarization on social media [28]–[34]. Random walk controversy [33] quantifies to what extent opinionated users are exposed to their own opinion rather than the opposite. Bubble radius [34] works on bipartite information networks and estimates the expected number of clicks to navigate from a page v to any page of opposing opinion. We focus on the metrics that better relate to our metric of *exposure to diverse information* (ExDIN). Differently from these two metrics, our measure of *exposure to diverse information* works on multi-partite information networks and quantifies users' exposure to diverse information click-after-click.

[2]https://www.mediawiki.org/wiki/ORES

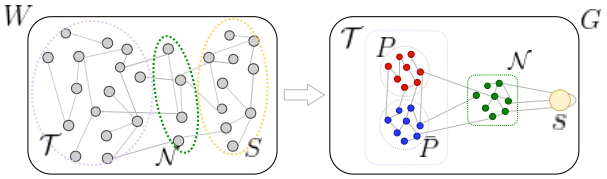


Fig. 1: On the left, the original Wikipedia graph. On the right, the final topic-induced network. The dashed circles in W are the set of nodes used to build the topic-induced network G. The colors red and blue refer to the sets P and $\bar{P}$, respectively. Green and yellow are $\mathcal{N}$ and s respectively. We keep the image tidy and do not specify the edges' direction.

Cultural bias on Wikipedia. Recent works found the presence of cultural bias in the same articles of different languages [35] and gender biases [36], [37]. These content-based analyses prove that Wikipedia can be subjected to bias. We decided to investigate bias on a novel topological perspective.

III. Preliminaries

We encode a topic into a topic-induced network, a subgraph of the entire English Wikipedia's graph $W = (A, L)$ (Fig. 1). The nodes of the graph are *articles* [38], and edges are links connecting pages, *wikilinks*.[3] Among all articles, we identify a set of pages $\mathcal{T} \subset A$, about the topic. We partition these pages into two sets P and $\bar{P}$ (i.e., $P \cap \bar{P} = \emptyset$ and $P \cup \bar{P} = \mathcal{T}$), each gathering articles about the same side of the topic. In addition to the articles in $\mathcal{T}$, we collect in $\mathcal{N}$ the pages at one-hop distance from them. In this way, we can consider the chances of moving across partitions via articles not necessarily related to the topic. To reduce the complexity of our analysis, we cluster all the pages in $A \setminus (\mathcal{T} \cup \mathcal{N})$, into one super node s. Note that s is only connected to vertices in $\mathcal{N}$. For each node $v \in \mathcal{N}$, we can have multiple edges going to s, which we compress into one. Respectively, s can have multiple links to node $v \in \mathcal{N}$, compressed into one as well.

A topic-induced network is the directed weighted graph $G = (V, E)$, whose set of articles V is $\mathcal{T} \cup \mathcal{N} \cup \{s\}$ of size $n + 1$, and edges E are the links connecting them. The edge weights are *transition probabilities* stored in a row-stochastic matrix $M \in [0, 1]^{|V|\times|V|}$, whose entry $m_{i,j}$ is the probability that from page i a reader moves to j, and it is set to 0 if $(i, j) \notin E$.

In practice, to build a topic-induced network, we first extract the entire Wikipedia's network from a complete English Wikipedia dump.[4] Then, we only retain the graph induced by $\mathcal{T}$, whose articles are selected and partitioned according to the strategy adopted in [27]. For instance, the topic *abortion* polarizes into *pro-life* (P) and *pro-choice* ($\bar{P}$) articles. The pro-life subcorpus consists of all articles categorized under

[3]We exclude links within the same page and resolve all the redirects [39]. We do consider links in the infoboxes, which are summary standardized tables at the top–right corner of articles.

[4]We refer to the dump of September 2020.

Topic	P	$\bar{P}$	Seed P	Seed $\bar{P}$
Abortion	Pro-life	Pro-choice	Anti-abortion movement	Abortion-rights movement
Cannabis	Prohibition	Activism	Cannabis prohibition	Cannabis activism
Guns	Control	Rights	Gun control advocacy groups	Gun rights advocacy groups
Evolution	Creationism	Evolutionary biology	Creationism	Evolutionary biology
Racism	Racism	Anti-racism	Racism	Anti-racism
LGBT	Discrimination	Support	Discrimination against LGBT people	LGBT rights movement

TABLE I: For each topic, the table indicates the partitions P and $\bar{P}$ to which each standing corresponds. Moreover, we report the seed category for each partition.

Topic	$\lvert V \setminus \{s\}\rvert$	$\lvert P\rvert$	$\lvert\bar{P}\rvert$	$\lvert\mathcal{N}\rvert$	$\lvert E\rvert$	$\lvert E_{P\to\bar{P}}\rvert$	$\lvert E_{\bar{P}\to P}\rvert$	$\lvert E_{\mathcal{N}\to P}\rvert$	$\lvert E_{\mathcal{N}\to\bar{P}}\rvert$
Abortion	56056	469	291	55296	2.1M	205	97	21396	29889
Cannabis	32743	45	231	32470	1.1M	8	6	656	27823
Guns	65743	167	187	65393	2.5M	98	115	56702	16608
Evolution	84788	342	1334	83113	1.99M	391	135	15601	58720
Racism	129963	1024	1022	127953	4.8M	746	560	74354	58195
LGBT	150563	459	640	149479	4.6M	195	143	92975	81706

TABLE II: Networks' statistics.

the *seed* category "Anti-abortion movement" and its subcategories. Similarly, we obtain the pro-choice corpus starting from the category "Abortion-rights movement." Because we want the partitions to be disjoint, articles belonging to both "Anti-abortion movement" and "Abortion-rights movement" are assigned to $\mathcal{N}$.

In fact, as a consequence of Wikipedia's Neutral Point of View (NPOV) policy [4], we assume articles' content to *"fairly and proportionately represent all the significant views that have been published by reliable sources on the topic."* Moreover, as subcategories are often redundant or not entirely related to the parent category, we check them manually, discarding categories whose names do not include topic-specific keywords.

A. Topic-Induced Networks

We collect the topic-induced networks related to six different polarizing topics: *abortion, cannabis, guns, evolution, LGBT*, and *racism* (Tab. I).

1) ***Partitions****:* In Tab. II,[5] we observe that the size of P and $\bar{P}$ differs substantially, for all the topics but *racism* and *guns*. The disproportionate number of articles does not imply an unbalance in content representation, but it can affect the partition's exposure within the entire Wikipedia network. The sizes of P and $\bar{P}$ are not linear in the number of edges across partitions. For instance, although the nodes in *pro-life* are twice as many as those in *pro-choice*, the links pointing to *pro-choice* are 36% more than those pointing to *pro-life*. This happens, with different magnitude, also for *guns* and *LGBT*.

2) ***Hyperlinks across partitions****:* The direct exposure of users in P to pages in $\bar{P}$, depends on the number of links

[5]We add to the set $\mathcal{N}$ the articles assigned to both partitions. The size of such intersections is: 2 (*abortion*), 3 (*cannabis*), 2 (*evolution*), 1 (*guns*), 5 (*LGBT*), 7 (*racism*). Because we do not remove these articles, they act as bridges connecting P and $\bar{P}$ in sessions longer than one click.

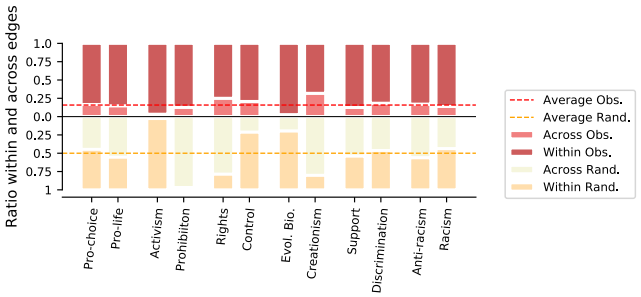


Fig. 2: Percentage of edges across and within partitions in each topic-induced network (red) and a random graph with the same degree distribution (orange). Topics in order are: *abortion*, *cannabis*, *guns*, *evolution*, *racism*, and *LGBT*.

p-values	$\mathcal{N}\to P$ vs. $\mathcal{N}\to\bar{P}$	Incoming higher avg.	$P\to\mathcal{N}$ vs. $\bar{P}\to\mathcal{N}$	Outgoing higher avg.
Abortion	$8.4\cdot10^{-2}$	-	$9.6\cdot10^{-1}$	-
Cannabis	$6.5\cdot10^{-8}$(**)	Activism	$1.6\cdot10^{-13}$(**)	Activism
Guns	$1.3\cdot10^{-4}$(**)	Control	$5.1\cdot10^{-2}$	-
Evolution	$7.2\cdot10^{-3}$(**)	Creationism	$4.9\cdot10^{-5}$(**)	Creationism
Racism	$6.2\cdot10^{-6}$(**)	Anti-racism	$3.3\cdot10^{-7}$(**)	Anti-racism
LGBT	$1.4\cdot10^{-2}$(**)	Discrimination	$1.4\cdot10^{-5}$(**)	Discrimination

TABLE III: We report the p-values of t-tests ($\alpha = 0.05$) on (1) the number of links from $\mathcal{N}$ to P and $\bar{P}$ (first column), and (2) the number of links to $\mathcal{N}$ from P and $\bar{P}$ (third column). In the second and fourth columns we indicate which partition is significantly more connected to the rest of Wikipedia. Statistics are computed after bootstrapping the distributions of from $\mathcal{N}$ to P and $\bar{P}$ and vice versa.

connecting the two partitions.[6] To study their connectivity, we compare the portion of links in pages of P pointing to P and $\bar{P}$, with the same quantities expected on a random graph with the same degree sequence. In Fig. 2, we observe that most of the hyperlinks point to pages of the same partition. On average fewer than 25% of links point toward the opposing partition, which is against the 50% expected on a random graph. The differences between the real and expected number of hyperlinks highlight that (1) links are, obviously, not randomly placed, (2) the strength of connections within and between partitions is skewed w.r.t. the distribution of edges conditioned on the number of nodes and their degree. Furthermore, we speculate that the higher number of hyperlinks directed to pages of the same partition is due to the intrinsic clustered nature of Wikipedia [40], [41].

3) ***Topic connectivity to the rest of Wikipedia****:* We briefly investigate the connectivity between P (resp. $\bar{P}$) with the rest of pages connected to it (i.e., $\mathcal{N}$). In Tab. III, we observe that for all the topics, but *abortion*, the average number of links coming from articles in $\mathcal{N}$ and pointing to articles in $\mathcal{N}$ is significantly higher for one of the two partitions.

4) ***Distribution of across-partition links****:* If across-partition links are uniformly placed within articles of a partition, users starting from an arbitrary node in the partition have the chance

[6]We note that the number of edges across *cannabis*'s partitions is low, nevertheless we keep the topic because on sessions longer that 1 click there are other paths connecting the partitions.

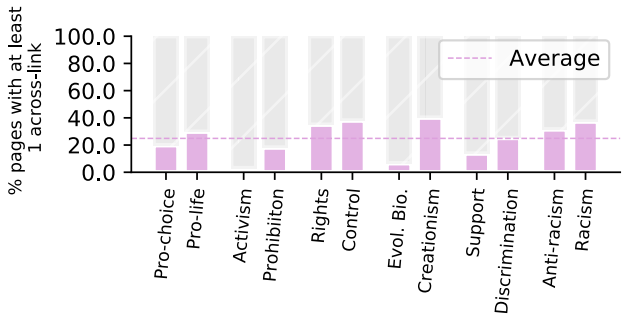


Fig. 3: Percentage of articles in P connected to $\bar{P}$. Topics are: *abortion*, *cannabis*, *guns*, *evolution*, *racism* and *LGBT*.

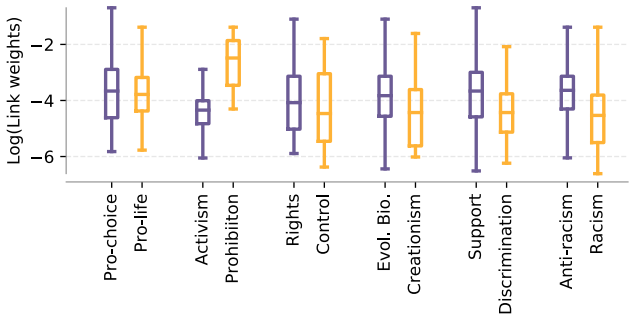


Fig. 4: Distributions of across-partition links weights. Topics are: *abortion*, *cannabis*, *guns*, *evolution*, *racism* and *LGBT*.

to visit pages about another branch of the topic. However, we observe that in our networks only a small subset of the pages expose their visitors to another branch of the topic: Fig. 3 shows that the average percentage of pages connecting to the other partition is 25% (average of 1.8 links per page), thus most of the nodes are not connected to the other partition.

5) ***Weight distribution of across-partition links****:* The likelihood of traversing a link connected to the other side is conditioned on the number of links in a page. Fig. 4 shows that, for each topic, there is one partition whose average probability of traversing an across-partition hyperlink is statistically higher than the other partition (according to *t-tests* with $\alpha = 0.05$). For instance, the average chance to go from *creationism* to *evolutionary biology* are significantly lower than moving in the opposite direction.

IV. Metrics

In this section we introduce the metrics to quantify the exposure to diversity, accounting for (1) the across-partition edges distribution over nodes, (2) the likelihood of traversing a link toward the other partition, and (3) the average exposure to diversity of all pages in a partition, considering navigation sessions of at least one click.

A. Model of Readers' Behavior

To comprehensively measure how much the network topology exposes users to diversity, we should consider both the graph topology and how readers navigate the network. Indeed, the exposure to diverse information might vary for users who behave differently in terms of navigation session length and next-link choices. So far, there are no models that generalize the navigation behavior of Wikipedia users. Thus, on top of previous findings [24], [25], in Sect. IV-A1 and IV-A2, we define a parametric model that simulates a wide range of users' navigation sessions by embedding different behaviors. We emphasize that the scope of this model is not to perfectly replicate users' behavior on Wikipedia. Rather, we want to see how users simulated from a reasonable and general model get exposed to diverse information.

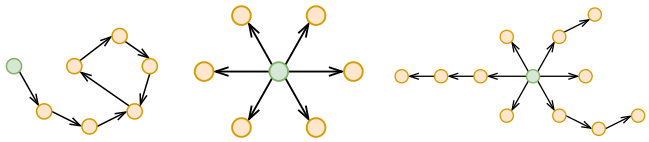
1) ***Model Clicks Within Pages (CwP)****:* When readers visit a page, they have the possibility of clicking any link. However, according to the information needs they want to satisfy, each of the links may have a different click-probability [7]. We characterize the probability of "clicking a link j within an article i" in three ways. First, let i be an article in V and $j \in N_{out}(i)$, where $N_{out}(i)$ is the the set of pages to which i has a link. We define $pos(j|i)$ as the rank of j among all links in i, and $r(j|i) = |N_{out}(i)| - pos(j|i)$, such that a higher value indicates a higher ranking position. We consider links in the infoboxes as at the top of the article, according to results in [24], [25]. Moreover, we introduce $\tanh x = \frac{e^{2x}-1}{e^{2x}+1}$, which we use to transform ranking positions to values between 0 and 1, such that links at the top of the page are assigned similar scores. For instance if two links are adjacent in a line, likely their probability of being clicked is similar.

We embed *clicks within pages* models (CwP) into G by setting its transition matrix M in one of the following modes:

1) M^u (*Uniform*), whose entry $m(i,j) = \frac{1}{|N_{out}(i)|}$ mimics readers who click each link uniformly at random;
2) M^p (*Position*), whose entry $m(i,j) = \frac{\tanh r(j|i)}{\sum_{j \in N_{out}(i)} \tanh r(j|i)}$ captures readers who click with higher probability links appearing on the top of the page. This model is based on previous works showing that the links' position is a good predictor to determine its success [18], [42];
3) M^c (*Clicks*), whose entry $m(i,j) = \frac{c_{i,j}}{\sum_{j \in N_{out}(i)} c_{i,j}}$ is the observed probability that users in i will click the link toward j. The quantity $c_{i,j}$ counts how many times on average real users clicked the hyperlink from page i to j, from August 2019 to September 2020.[7] For the links never clicked, we set $c_{ij} = 10$, the minimum number of times that the link must be clicked to be included in the dataset [44]. This smoothing factor allows one to assign a positive weight to links rarely clicked.

2) ***Readers Navigation Model****:* To characterize the users' sessions, we define a stochastic process with $|V|$ states, which,

[7]Wikipedia's clickstream data is publicly available and preserves users' privacy [15], [43]. Data description at Research:Wikipedia_clickstream.



(a) Random Navigation ($\alpha = 0$) (b) Star-like ($\alpha = 1$) (c) Star-like Random Navigation ($0 < \alpha < 1$)

Fig. 5: Navigation model for different α.

for each click, approximates the probability of reaching any of the articles starting at random from $p \in P$ (or from $\bar{P}$). We consider the process $\{X^\ell; \ell = 0, 1, \dots L\}$, on the set of nodes V induced by the transition matrix M with starting state X^0 selected from the probability distribution $\boldsymbol{\pi}_P^0 = \{(\pi_P)_i\} \in \mathbb{R}^{1\times n}$ over V. Assuming that the user session length (the number of clicks) is finite, we evaluate the process on a finite number of steps L. Thus, $\mathbf{Pr}(X^\ell = j) = (\pi_P^\ell)_j$, where the (row) vector $\boldsymbol{\pi}_P^\ell$ is given by the following variation of the Personalized Random Walk with Restart (RWR).

Definition 1 (Navigation Model): Let M_0 be the transition matrix embedding a click-within-pages model, $\boldsymbol{\pi}_P^0$ the distribution of the starting state over P, and $\alpha \in [0, 1]$ the restart parameter. We have

$$\boldsymbol{\pi}_P^1 = \boldsymbol{\pi}_P^0 \cdot M_0 \tag{1}$$

and, for $\ell \geq 1$,

$$\boldsymbol{\pi}_P^{\ell+1} = (1-\alpha)\boldsymbol{\pi}_P^\ell \cdot M_\ell + \alpha(\boldsymbol{\pi}_P^0 \cdot M_\ell), \tag{2}$$

where $M_\ell = norm((D(M_{\ell-1})^T)^T)$ and $D = diag\left(\mathbf{1} + \boldsymbol{\pi}_P^{\ell-1}\right)^{-1}$. The operator $norm(M)$ transforms matrix M into a right-stochastic matrix by normalizing each row independently such that it sums to 1.

This process is a variation of the standard random-surfer model that differs for the update of the transition matrix at each step. The vector $\boldsymbol{\pi}_P^\ell$ represents the likelihood that each node is reached at step ℓ if the session starts uniformly at random from a node in P. Assuming that readers do not click multiple times the same link within a session, we desire to deflate the probability of reaching nodes that, at step $\ell + 1$, have already been visited with high probability. We achieve this by dividing the rows of M by the vector of probabilities $\boldsymbol{\pi}_P^\ell$+1, where 1 is a smoothing factor, and then normalize the matrix to obtain the updated stochastic matrix to use in the next iteration. Looking deeper into the model:

- For $\alpha = 0$ (Fig. 5(a)), the readers' clicks depend only on the CwP model. In this case, especially if related articles are not densely connected, the exploration can quickly lead to articles less related to the starting page.
- For $\alpha = 1$ (Fig. 5(b)), readers locally explore articles likely semantically related to each other [1], and the model emulates a *star-like* behavior, which consists in sequentially opening links from the starting page.
- For $0 < \alpha < 1$ (Fig. 5(c)), the readers' choices depend on the CwP model and, occasionally, they go back to the initial page. The more α is close to 1 the more users show a star-like behavior. The closer α is to 0 the more users navigate navigate in a more Depth First Search-oriented fashion. The model emulates (1) readers who sequentially explore articles and then jump back to the starting page, or (2) readers keeping open multiple paths.

Wikipedia does not have a button that allows readers to go back to the previous page. Thus, to *jump back* consists of clicking the browser's back button, until the session starting page. The restart parameter indirectly embeds the back button, which for the absence of *back-links* on Wikipedia does not appear in the graph.

B. Quantification of Exposure to Diverse Information

The *exposure to diverse information* aims to quantify how much the network structure allows readers to reach one, or multiple sets of articles, depending on their behavior. It is built upon both the CwP and the *navigation* models, and its application generalizes to arbitrary sets of nodes in a graph.

Definition 2 (Exposure to diverse information (ExDIN)): Given two sets of pages $P, \bar{P}$ in V, let $\boldsymbol{\pi}_P^\ell$ be the vector indicating the probability distribution of reaching any node in V at step ℓ ($\ell \geq 1$) starting from a random page in P. We say that the exposure of P to $\bar{P}$ is

$$e_{P\to\bar{P}}^\ell = \sum_{j\in\bar{P}} \mathbf{Pr}(X^\ell = j) = \sum_{j\in\bar{P}} (\pi_P^\ell)_j \tag{3}$$

and it describes the probability that a reader in P reaches an arbitrary node in $\bar{P}$ at the ℓth click.

Definition 2 can be extended to multiple sets. Assume that we want to measure how much the set P is exposed to three sets of nodes, Q, Z, and L. The total exposure to the three sets is the ExDIN computed setting $\bar{P} = Q \cup Z \cup L$. Otherwise, if we want to have the ExDIN w.r.t. to each set, namely, $e_{P\to Q}, e_{P\to Z}$, and $e_{P\to L}$, we take $\boldsymbol{\pi}_P^\ell$ and sum up the probabilities of the nodes within each set.

To quantify the extent to which the exposure to diverse information is balanced across P and $\bar{P}$, we introduce the *mutual exposure to diverse information*.

Definition 3 (Mutual exposure to diverse information (M-ExDIN)): Let $e_{P\to\bar{P}}^\ell$ and $e_{\bar{P}\to P}^\ell$ be the exposure to diverse information of sets P and $\bar{P}$. The mutual exposure between the sets is

$$\epsilon^\ell = \frac{\min\{e_{P\to\bar{P}}^\ell, e_{\bar{P}\to P}^\ell\}}{\max\{e_{P\to\bar{P}}^\ell, e_{\bar{P}\to P}^\ell\}} \in [0, 1]. \tag{4}$$

If either $e_{P\to\bar{P}}^\ell$ or $e_{\bar{P}\to P}^\ell$ is 0, then $\epsilon = 0$.

The closer ϵ is to 1, the more balanced are the probabilities of moving from one set to the other are. Thus, the network topology does not favor connections from one set to the other. Otherwise, the network structure tends to favor the navigation from one partition toward the other. With this view, if the network structure facilitates to move from one of the sets to the other, we may say that the network topology is *biased* toward a direction. Thus, M-ExDIN is a measure of the network's bias w.r.t. two sets of nodes, at each click of a session. If

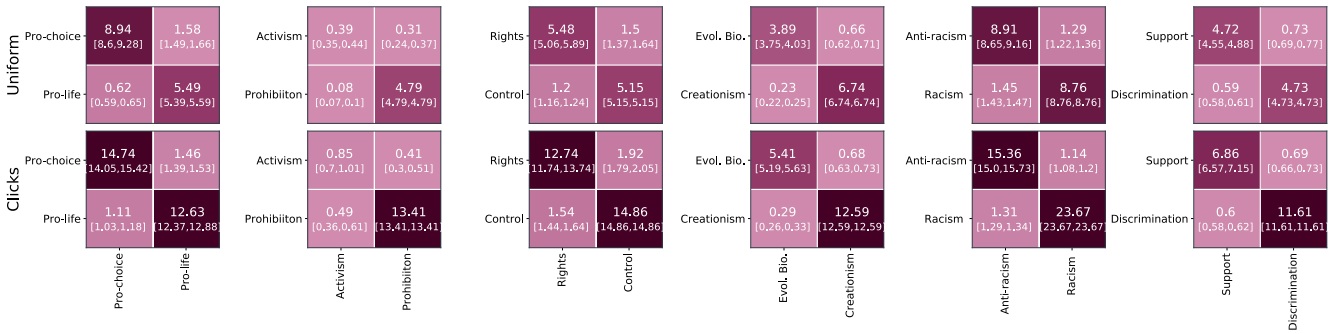


Fig. 6: *Local exposure to diversity*. Each plot shows the adjusted-ExDIN (%) across partitions. On the main diagonal we have flows within the same partition, e.g., $P \to P$. The off-diagonal reports the probability of moving across sides, e.g., $P \to \bar{P}$. The y-axis indicates the source and the x-axis is the destination. To each row corresponds the exposure to diverse information computed for different CwP. Darker colors indicate higher probability of being in the corresponding square in one click. The values in the brackets are the 90% confidence intervals. Topics are: *abortion, cannabis, guns, evolution, racism*, and *LGBT*.

the sizes of sets P and $\bar{P}$ are unbalanced, we obtain higher probabilities for large partitions. To take the sizes into account, we introduce a strategy to compute the adjusted-ExDIN. Given the two partitions P and $\bar{P}$, we set a sample size equal to $z = \min\{|P|, |\bar{P}|\}$. From the two sets we sample with replacement P'_i and $\bar{P}'_i$ of size z, respectively from the initial partitions. Hence, we bootstrap $e^{\ell}_{P' \to \bar{P}'}$ estimating the value of the adjusted-ExDIN.

V. Exposure to Diverse Viewpoints

A. Local Exposure to Diversity

The *local* exposure to diversity is the possibility of accessing articles about another branch of the topic within one click. We measure it using exposure to diverse information, setting $\ell = 1$, which describes the static connectivity among partitions accounting for users' choices and the network topology.

In Fig. 6, the local exposure to diversity on each topic-induced network shows the following:[8] When considering only the graph topology (i.e., M^u): (1) The networks' topology facilitates users to remain in a *knowledge bubble* (i.e., same partition) hindering the exploration of the topic's diverse stances. For every topic the probability of visiting pages of the opposing partition is on average 12 times lower than staying in the initial one. (2) One of the two partitions induces higher chances of remaining within the same bubble. On average, among all topics but *cannabis*, one of the two partitions has 2 times more chances of keeping users within its articles.

After embedding users' past behavior (i.e., M^u): (3) The probability that readers keep visiting pages about the same topic click-after-click (i.e., remaining within a *knowledge bubble*) is higher than when users click links uniformly at random (i.e., M^u). Indeed, on average, users have 1.5 more chances of moving within $\mathcal{T}$. The probability increases significantly, suggesting to start a discussion on the importance of exposing users to diverse information. (4) The likelihood of exploring diverse content slightly increases, showing that real users clicked pages of the opposing partitions more than what described by the uniform model. On average, users have 1.6 times more chances of moving to the opposing sides. (5) The discrepancy between the probability of moving within and outside the initial partition is on average even 2.6 times larger than when using M^u.

From past users' clicks we observe the preference of navigating across pages of the same partition. So far, users' behavior has always been justified by their information needs. From the observations (1) and (2), we know that the network's topology potentially favors the visit of pages of same standing. Is it possible that users' next-click choices are influenced by the hyperlinks network?

B. Dynamic Exposure to Diversity

We expand the analysis to sessions longer than one click to study the users' dynamic exposure to diversity. In Fig. 7, we observe that: (1) Within a few steps (4–5 clicks), users are more exposed to the pages of the same partition (i.e., the ratio of exposures is smaller than 0). (2) The current structure of the network favors users starting from any random page to reach one partition more easily than the other. Can we consider it a bias of the network? Is it fair that one side of a polarizing topic is less reachable than the opposite from any random node in the graph? (3) If users navigate according to a star-like random navigation model, the ratio between moving outside and within the partition stays steady or slightly increases. (4) All topic-induced networks are topologically biased; indeed none of them does the network provides an equal exposure across partitions (i.e., mutual exposure to diverse information is lower than 100%). (5) The local bias is smaller than the overall bias of the network. In general, the mutual exposure to diverse information decreases for longer sessions. (6) The

[8]We omit the plots showing result when using the model embedding the links' position. In general, there are not significant differences compared to the uniform model. For some topics, such as *guns*, the links' position plays a more significant role, worsening the user exposure to diverse information.

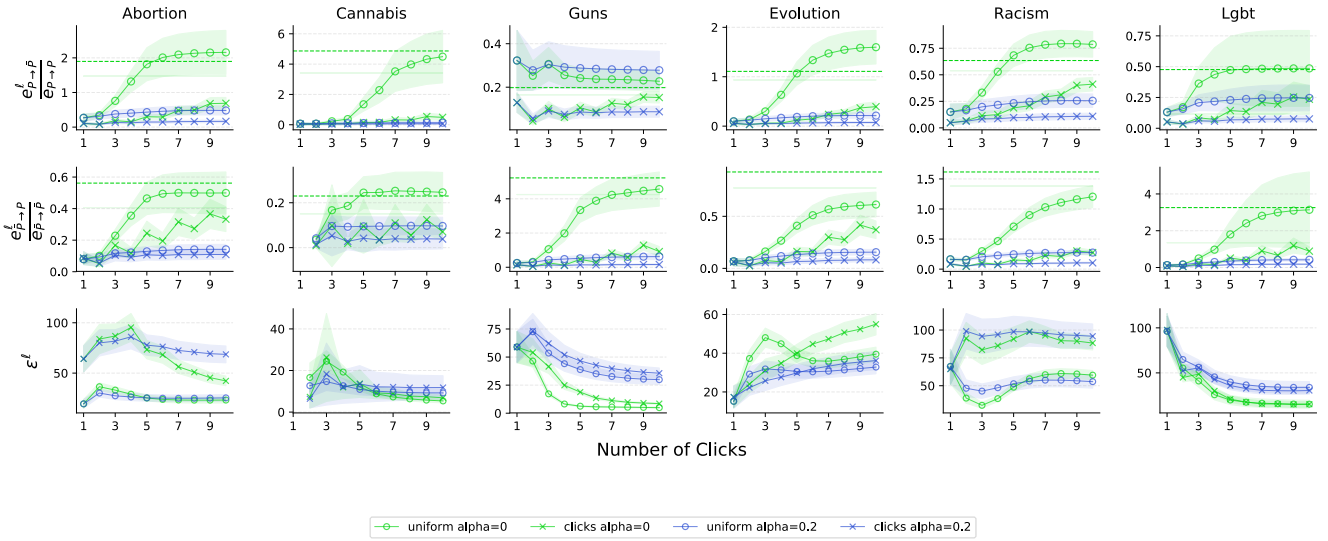


Fig. 7: *(Adjusted) Dynamic Exposure to Diversity for sessions of 1 to 10 clicks*. The x-axis indicates the number of clicks. In the first two rows, the y-axis is the ratio of the probability of moving from P to $\bar{P}$ and P, and vice versa. Values lower than 1 indicate that remaining in the *knowledge bubble* is more likely than visiting pages of diverse content. Values close to zero quantify the chances of staying in a *knowledge bubble*. The y-axis of the third row indicates the mutual exposure to diversity (%). Values closer to 0 show that the current network topology might be biased toward the partition P or $\bar{P}$. Colors indicate the navigation model (set by α). Shapes encode the CwP model. The bands indicate the standard deviation. The dashed lines represent the convergence values for infinite sessions.

general topology of the graph makes pages related to *liberal* standings more accessible.[9]

The lack of mutual exposure might depend on many factors such as *underlinked* articles [1] or missing words to attach links. An in-depth investigation of this condition may be an interesting future work.

C. Factors Related to Exposure to Diversity

1) ExDIN and homophily: We measure the Pearson correlation between the homophily[10] of article i and the users' exposure to diversity starting a session in i. Locally, starting from a page with high homophily decreases users' probabilities of being exposed to diverse content. Indeed, we observe a negative correlation for 1–3 clicks sessions (avg. -0.45), and an average correlation close to 0, for longer sessions.

2) ExDIN and centrality: We measure the Pearson correlation between the degree centrality of article i and the users' exposure to diversity starting a session in i. The number of incoming links of an initial page does not play a role in determining the exposure to diversity, especially within a few clicks (in-degree correlation 0.07). On the other hand, within 1–3 clicks, the lower out-degree mildly increases the users exposure to diversity (average -0.25). Under the uniform model, it is because the links' transition probabilities of pages with a few links are larger than pages with a lot of links.

VI. Discussions

In this work, we look for the first time at Wikipedia's hyperlink structure to measure its influence on users' exposure to diverse information. By employing two Wikipedia-tailored metrics, we quantify the likelihood of visiting pages representing different aspects of a topic throughout a navigation session. Our findings indicate that the current network topology often limits exposure to diverse information and incentivizes users to remain in *knowledge bubbles*.

The ultimate goal of this work is to draw attention and initiate a discussion about the importance of evaluating the hyperlink structure as part of Wikipedia's goal to provide a natural point of view presentation, even for polarizing subjects. Our observations raise a number of interesting questions for the Wikipedia community. As an example, consider a page about an *anti-abortion organization*. It seems natural that this page has more hyperlinks to pages related to anti-abortion subjects than to pages related to abortion rights. This is reasonable and aligns with the current purpose of Wikipedia's internal links, but is it still reasonable, and conforms with the goal of a natural point of view? Similarly, we observe that in the directed hyperlink graph, it is often more likely to reach an article about B starting from an article about A, than reaching an article about A starting with an article about B, when A and

[9]We omit the star-like model ($\alpha = 0$), because its value is steady around the value for $\ell = 1$, and we omit the position CwP model because it shows trends similar to uniform.

[10]We use the EI Homophily index [45]: $EI(v \in P) = \frac{|ext_P| - |int_P|}{|ext_P| + |int_P|}$, where ext_P is the set of edges from P to the rest of the network, and int_P is the number of edges pointing to P.

B represent two aspects of a topics. Again, some imbalance is reasonable, but it can keep users locked in an information bubble. How do we distinguish between the two cases?

We expect ours and future findings to motivate work on editors' support tools for contextualizing pages within their neighborhood in the hyperlink network, and suggest hyperlink modifications to improve access to diverse content.

ACKNOWLEDGEMENTS

The project is supported by DARPA LwLL program, by the ERC Advanced Grant 788893 AMDROMA, the EC H2020RIA project "SoBigData++" (871042), and the MIUR PRIN project ALGADIMAR. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[1] Wikipedia, "Linking," in *Wikipedia:Manual of style/linking*.
[2] B. Keegan, D. Gergle, and N. Contractor, "Hot off the wiki: dynamics, practices, and structures in wikipedia's coverage of the tōhoku catastrophes," in *Proc. of the 7th international symposium on Wikis and open collaboration*, 2011.
[3] A. Piscopo and E. Simperl, "What we talk about when we talk about wikidata quality: a literature survey," in *Proc. of the 15th International Symposium on Open Collaboration*, 2019.
[4] Wikipedia, "Neutral point of view," in *Wikipedia:Neutral_point_of_view*.
[5] I. Beschastnikh, T. Kriplean, and D. W. McDonald, "Wikipedian self-sovernance in action: motivating the policy lens." in *ICWSM*, 2008.
[6] A. Forte, V. Larco, and A. Bruckman, "Decentralization in wikipedia governance," *Journal of Management Information Systems*, 2009.
[7] P. Singer, F. Lemmerich, R. West, L. Zia, E. Wulczyn, M. Strohmaier, and J. Leskovec, "Why we read wikipedia," in *Proc. of the 26th International Conference on World Wide Web*, 2017.
[8] Wikipedia, "Statistics," in *Wikipedia:Statistics*.
[9] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proc. of FAccT 2020*.
[10] M. Redi, B. Fetahu, J. Morgan, and D. Taraborelli, "Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability," in *The World Wide Web Conference*, 2019.
[11] B. Fetahu, K. Markert, W. Nejdl, and A. Anand, "Finding news citations for wikipedia," in *Proc. of the 25th ACM International on Conference on Information and Knowledge Management*, 2016.
[12] T. Piccardi, M. Catasta, L. Zia, and R. West, "Structuring wikipedia articles with section recommendations," in *The 41st International ACM SIGIR Conference*, 2018.
[13] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proc. of the 25th International Conference on World Wide Web*, 2016.
[14] A. Paranjape, R. West, L. Zia, and J. Leskovec, "Improving website hyperlink structure using server logs," in *Proc. of the Ninth ACM International Conference on Web Search and Data Mining*, 2016.
[15] E. Wulczyn, R. West, L. Zia, and J. Leskovec, "Growing wikipedia across languages via recommendation," in *Proc. of the 25th International Conference on World Wide Web*, 2016.
[16] D. Helic, M. Strohmaier, M. Granitzer, and R. Scherer, "Models of human navigation in information networks based on decentralized search," in *Proc. of the 24th ACM conference on hypertext and social media*, 2013.
[17] P. Gildersleve and T. Yasseri, "Inspiration, captivation, and misdirection: Emergent properties in networks of online navigation," in *International Workshop on Complex Networks*, 2018.
[18] D. Lamprecht, K. Lerman, D. Helic, and M. Strohmaier, "How the structure of wikipedia articles influences user navigation," *New Review of Hypermedia and Multimedia*, 2017.
[19] P. Singer, T. Niebler, M. Strohmaier, and A. Hotho, "Computing semantic relatedness from human navigational paths: A case study on wikipedia," in *International Journal on Semantic Web and Information Systems 9*, 2013.
[20] R. West and J. Leskovec, "Human wayfinding in information networks," in *Proc. of the 21st international conference on World Wide Web*, 2012.
[21] A. T. Scaria, R. M. Philip, R. West, and J. Leskovec, "The last click: Why users give up information network navigation," in *Proc. of the 7th ACM international conference on Web search and data mining*, 2014.
[22] A. Dallmann, T. Niebler, F. Lemmerich, and A. Hotho, "Extracting semantics from random walks on wikipedia: Comparing learning and counting methods." in *Wiki@ICWSM*, 2016.
[23] T. Koopmann, A. Dallmann, L. Hettinger, T. Niebler, and A. Hotho, "On the right track! analysing and predicting navigation success in wikipedia," in *Proc. of the 30th ACM Conference on Hypertext and Social Media*, 2019.
[24] D. Dimitrov, P. Singer, F. Lemmerich, and M. Strohmaier, "Visual positions of links and clicks on wikipedia," in *Proc. of the 25th International Conference Companion on World Wide Web*, 2016.
[25] ——, "What makes a link successful on wikipedia?" in *Proc. of WWW 2017*.
[26] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE signal processing magazine*, 2010.
[27] F. Shi, M. Teplitskiy, E. Duede, and J. A. Evans, "The wisdom of polarized crowds," *Nature human behaviour*, 2019.
[28] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proc. of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.
[29] A. Cossard, G. De Francisci Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini, "Falling into the echo chamber: The Italian vaccination debate on Twitter," in *Proc. of the International AAAI Conference on Web and Social Media*, 2020.
[30] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Fifth international AAAI conference on weblogs and social media*, 2011.
[31] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public opinion quarterly*, 2016.
[32] P. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in *7th International AAAI Conference on Weblogs and Social Media*, 2013.
[33] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *ACM Transactions on Social Computing*, 2018.
[34] S. Haddadan, C. Menghini, M. Riondato, and E. Upfal, "Repbublik: Reducing polarized bubble radius with link insertions," in *Proc. of WSDM 2021*.
[35] E. S. Callahan and S. C. Herring, "Cultural bias in wikipedia content on famous persons," *Journal of the American society for information science and technology*, 2011.
[36] E. Graells-Garrido, M. Lalmas, and F. Menczer, "First women, second sex: Gender bias in wikipedia," in *Proc. of the 26th ACM Conference on Hypertext & Social Media*, 2015.
[37] C. Wagner, E. Graells-Garrido, D. Garcia, and F. Menczer, "Women through the glass ceiling: gender asymmetries in wikipedia," *EPJ Data Science*, 2016.
[38] Wikipedia, "Namespace," in *Wikipedia:Namespace*.
[39] ——, "Redirect," in *Wikipedia:Redirect*.
[40] U. Brandes, P. Kenis, J. Lerner, and D. Van Raaij, "Network analysis of collaboration structure in wikipedia," in *Proc. of the 18th international conference on World wide web*, 2009.
[41] D. Lizorkin, O. Medelyan, and M. Grineva, "Analysis of community structure in wikipedia," in *International conference on World wide web*, 2009.
[42] D. Dimitrov, P. Singer, F. Lemmerich, and M. Strohmaier, "What makes a link successful on wikipedia?" in *Proc. of the 26th International Conference on World Wide Web*, 2017.
[43] D. Dimitrov and F. Lemmerich, "Democracy and difference: Different topic, different traffic: How search and navigation interplay on wikipedia," *The Journal of Web Science 6*, 2019.
[44] E. Wulczyn and D. Taraborelli, "Wikipedia clickstream," *https://doi.org/10.6084/m9.figshare.1305770.v22*, 2017.
[45] D. Krackhardt and R. N. Stern, "Informal networks and organizational crises: An experimental simulation," *Social Psychology Quarterly*, 1988.