**Noname manuscript No.**
(will be inserted by the editor)

# Misleading Stars: What Cannot Be Measured in the Internet?

**Yvonne-Anne Pignolet · Stefan Schmid · Gilles Tredan**

**Abstract** Traceroute measurements are one of the main instruments to shed light onto the structure and properties of today's complex networks such as the Internet. This article studies the feasibility and infeasibility of inferring the network topology given traceroute data from a worst-case perspective, i.e., without any probabilistic assumptions on, e.g., the nodes' degree distribution. We attend to a scenario where some of the routers are anonymous, and propose two fundamental axioms that model two basic assumptions on the traceroute data: (1) each trace corresponds to a real path in the network, and (2) the routing paths are at most a factor $1/\alpha$ off the shortest paths, for some parameter $\alpha \in (0, 1]$. In contrast to existing literature that focuses on the cardinality of the set of (often only minimal) inferrable topologies, we argue that a large number of possible topologies alone is often unproblematic, as long as the networks have a similar structure. We hence seek to characterize the set of topologies inferred with our axioms. We introduce the notion of star graphs whose colorings capture the differences among inferred topologies; it also allows us to construct inferred topologies explicitly. We find that in general, inferrable topologies can differ significantly in many important aspects, such as the nodes' distances or the number of triangles. These negative results are complemented by a discussion of a scenario where the trace set is best possible, i.e., "complete". It turns out that while some properties such as the node degrees are still hard to measure, a complete trace set can help to determine global properties such as the connectivity.

Y.-A. Pignolet
ABB Research, Switzerland
E-mail: yvonne-anne.pignolet@ch.abb.com

S. Schmid
Telekom Innovation Laboratories & TU Berlin, Germany
E-mail: stefan@net.t-labs.tu-berlin.de

G. Tredan
LAAS-CNRS, France
E-mail: gilles.tredan@laas.fr

## 1 Introduction

Surprisingly little is known about the structure of many important complex networks such as the Internet. One reason is the inherent difficulty of performing accurate, large-scale and preferably synchronous measurements from a large number of different vantage points. Another reason are privacy and information hiding issues: for example, network providers may seek to hide the details of their infrastructure to avoid tailored attacks.

Knowledge of the network characteristics is crucial for many applications as well as for an efficient operation of the network (e.g., for traffic engineering purposes [27], multipath TCP applications [19], reliable multicast [22], or ISP-enhanced CDN assignments [25], to just give some examples). Consequently, the research community has implemented many measurement tools to analyze at least the main properties of the network. Results of such measurements can then, e.g., be used to design more efficient network protocols in the future.

This article focuses on the most basic characteristic of the network: its *topology*. The classic tool to study topological properties is *traceroute*. Traceroute allows us to collect traces from a given source node to a set of specified destination nodes. A trace between two nodes contains a sequence of identifiers describing a route between source and destination. However, not every node along such a path is configured to answer with its identifier. Rather, some nodes may be *anonymous* in the sense that they appear as stars ('$*$') in a trace. Anonymous nodes exacerbate the exploration of a topology because already a small number of anonymous nodes may increase the spectrum of inferrable topologies that correspond to a trace set $\mathcal{T}$.

This article is motivated by the observation that the mere number of inferrable topologies alone does not contradict the usefulness or feasibility of topology inference; if the set of inferrable topologies is homogeneous in the sense that the different topologies share many important properties, the generation of all possible graphs can be avoided: an arbitrary representative may characterize the underlying network accurately. Therefore, we identify important topological metrics such as diameter or maximal node degree and examine how "close" the possible inferred topologies are with respect to these metrics.

## 1.1 Related Work

Arguably one of the most influential measurement studies on the Internet topology was conducted by the Faloutsos brothers [14] who show that the Internet exhibits a skewed structure: the nodes' out-degree follows a power-law distribution. Moreover, this property seems to be invariant over time. These results complement discoveries of similar distributions of communication traffic which is often self-similar, and of the topologies of natural networks such as human respiratory systems. The power-law property allows us to give good predictions not only on node degree distributions but also, e.g., on the expected number of nodes at a given hop-distance. Since [14] was published, many additional results have been obtained, e.g., [4,5,13,21], also by conducting a distributed computing approach to increase the number of measurement points [11]. However, our understanding of the Internet's structure remains preliminary, and the topic continues to attract much attention from the scientific communities. In contrast to these measurement studies, we pursue a more formal approach, and a complete review of the empirical results obtained over the last years is beyond the scope of this article.

The tool traceroute has been developed to investigate the *routing behavior* on the Internet. [11,15,20,23,26] Paxson [23] used traceroute to analyze pathological conditions, routing stability, and routing symmetry. To give another example, Gill et al. [15] demonstrate that large content providers (e.g., Google, Microsoft, Yahoo!) are deploying their own wide-area networks, bringing their networks closer to users, and bypassing Tier-1 ISPs on many routing paths.

Traceroute is also used to discover Internet *topologies* [12]. Unfortunately, there are several problems with this approach that render topology inference difficult, such as *aliasing* (a node appears with different identifiers on different interfaces) or *load-balancing*, which has motivated researchers to develop new tools such as *Paris Traceroute* [8, 18].

There are several results on traceroute sampling and its limitations. In their famous papers, Lakhina et al. [21] (by simulation) and Achlioptas et al. [4] (mathematically) have shown that the degree distribution under traceroute sampling exhibits a power law even if the underlying degree distribution is Poisson. Dall'Asta et al. [13] show that the edge and vertex detection probability depends on the betweenness centrality of each element and they propose improved mapping strategies. Another interesting result is due to Barford et al. [9] who experimentally show that when performing traceroute-like methods to infer network topologies, it is more useful to increase the number of destinations than the number of trace sources.

One drawback of using traceroute to determine the network topology stems from the fact that routers may appear as stars (i.e., anonymous nodes) in the trace since they are overloaded or since they are configured not to send out any ICMP responses. [28] The lack of complete information in the trace set renders the accurate characterization of Internet topologies difficult.

This article attends to the problem of what information on the underlying topology can be inferred despite anonymous nodes and assumes a conservative, "worst-case" perspective that does not rely on any assumptions on the underlying network.

An early work on this subject is by Yao et al. [28] who study possible candidate topologies for a given trace set and compute the *minimal topology*, that is, the topology with the minimal number of anonymous nodes. Answering this question turns out to be NP-hard. Subsequently, different heuristics addressing this problem have been proposed [16,18].

Our work is motivated by a series of papers by Acharya and Gouda. In [3], a network tracing theory model is introduced where nodes are "irregular" in the sense that each node appears in at least one trace with its real identifier. In [1], hardness results on finding the minimal topology are derived for this model. However, as pointed out by the authors themselves, the irregular node model—where nodes are anonymous due to high loads—is less relevant in practice and hence they consider strictly anonymous nodes in their follow-up studies [2]. Moreover, as proved in [2], the problem is still hard (in the sense that there are many minimal networks corresponding to a trace set), even for networks with only two anonymous nodes, symmetric routing and without aliasing.

In contrast to the line of research on cardinalities of minimal networks, we are interested in the *network properties* of inferrable topologies. If the inferred topologies share the most important characteristics, the negative results on cardinalities in [1,2] may be of little concern. Moreover, we believe that a study limited to minimal topologies only may miss important redundancy aspects of the Internet. Unlike [1,2], our work is constructive in the sense that algorithms can be derived to compute inferred topologies.

More remotely, our work is also related to the field of *network tomography*, where topologies are explored using pairwise end-to-end measurements, without the cooperation of nodes along these paths. This approach is quite flexible and applicable in various contexts, e.g., in social networks. For a good discussion of this approach as well as results for a routing model along shortest and second shortest paths see [7]. For example, [7] shows that for sparse random graphs, a relatively small number of cooperating participants is sufficient to determine the network topology fairly well.

Finally, researchers have recently also been interested in topology inference by *graph or virtual network embeddings* [24] (rather than exploration of single paths), as well as in the orthogonal question of inferring *capacities* or weights on a topology (see, e.g., [10, 17], and [6] for a result on multi-agent exploration). .

### 1.2 Our Contribution

This article initiates the study and characterization of topologies that can be inferred from a given trace set computed with the traceroute tool. While existing literature assuming a worst-case perspective has mainly focused on the cardinality of minimal topologies, we go one step further and examine specific topological graph properties.

We introduce a formal theory of topology inference by proposing basic axioms (i.e., assumptions on the trace set) that are used to guide the inference process. We present a novel definition for the isomorphism of inferred topologies which is aware of traffic paths; it is motivated by the observation that although two topologies look equivalent up to a renaming of anonymous nodes, the same trace set may result in different paths. Moreover, we propose the study of two extremes: in the first scenario, we only require that each link appears at least once in the trace set; interestingly, however, it turns out that this is often not sufficient, and we propose a "best case" scenario where the trace set is, in some sense, *complete*: it contains paths between all pairs of non-anonymous nodes.

The main result of the article is a negative one. It is shown that already a small number of anonymous nodes in the network renders topology inference difficult. In particular, we prove that in general, the possible inferrable topologies differ in many crucial aspects, e.g., the maximal node degree, the diameter, the stretch, the number of triangles and the number of connected components.

We introduce the concept of the *star graph* of a trace set that is useful for the characterization of inferred topologies. In particular, colorings of the star graphs allow us to constructively derive inferred topologies. (Although the general problem of computing the set of inferrable topologies is related to NP-hard problems such as *minimal graph coloring* and *graph isomorphism*, some important instances of inferrable topologies can be computed efficiently.) The chromatic number (i.e., the number of colors in the minimal proper coloring) of the star graph defines a lower bound on the number of anonymous nodes from which the stars in the traces could originate from. And the number of possible colorings of the star graph—a function of the *chromatic polynomial* of the star graph—gives an upper bound on the number of inferrable topologies. We show that this bound is tight in the sense that trace sets with that many inferrable topologies indeed exist. In particular, there are problem instances where the cardinality of the set of inferrable topologies equals the *Bell number*. This insight complements existing cardinality results and generalizes topology inference to arbitrary, not only minimal, inferrable topologies.

Finally, we examine the scenario of *fully explored networks* for which "complete" trace sets are available. As expected, the inferrable topologies are more homogenous in this case and can be characterized well with respect to many properties such as the distances between nodes. However, we also find that some other properties are inherently difficult to estimate. Interestingly, our results indicate that full exploration is often useful to derive bounds on global properties (such as connectivity) while it does not help much for bounds on more local properties (such as node degrees).

### 1.3 Organization

The remainder of this article is organized as follows. Our axioms for topology inference and some definitions are introduced in Section 2. Our main contributions are presented in Sections 3 and 4 where we derive bounds on inferrable topologies for general trace sets and fully explored networks, respectively. In Section 5, the article concludes with a discussion of our results and directions for future research.

## 2 Model

Let $\mathcal{T}$ denote the set of traces obtained from probing (e.g., by traceroute) a network $G_0 = (V_0, E_0)$ with *nodes* or *vertices* $V_0$ (the set of routers) and undirected *links* or *edges* $E_0$. We assume that $G_0$ is static during the probing time (or that probing is instantaneous), but we do not required that $G_0$ is connected. Each trace $T(u, v) \in \mathcal{T}$ describes a path connecting two nodes $u, v \in V_0$; when $u$ and $v$ do not matter or are clear from the context, we simply write $T$. Moreover, let $d_T(u, v)$ denote the distance (number of hops) between two nodes $u$ and $v$ in trace $T$. We define $d_{G_0}(u, v)$ to be the corresponding shortest path distance in $G_0$. Note that a trace between two nodes $u$ and $v$ may not describe the shortest path between $u$ and $v$ in $G_0$.

The nodes in $V_0$ fall into two categories: *anonymous* nodes and *non-anonymous* (or shorter: *named*) nodes. As it is the case in most related literature, We assume these categories to be permanent over time and traces: a node is either consistently anonymous or consistently non-anonymous. This is motivated by the fact that while sometimes nodes can be anonymous due to temporary events such as high loads, most of the time the cause are static configurations, see [28]: Some routers are configured to not send out ICMP responses while others use the destination addresses of traceroute packets instead of their own addresses as source addresses for outgoing ICMPv6 packets. In both cases, the presence of a router, but not its address, can be detected by traceroute. Also, IPv4 and IPv6 routers with ICMP disabled, as well as IPv6 routers not configured with global addresses, appear as anonymous. Such router configurations exist for numerous reasons, e.g., administrative overhead or security reasons.

Each trace $T \in \mathcal{T}$ hence describes a sequence of symbols representing anonymous and non-anonymous nodes. We make the natural assumption that the first and the last node in each trace $T$ is non-anonymous. Moreover, we assume that traces are given in a form where non-anonymous nodes appear with a unique, anti-aliased identifier (i.e., the multiple IP addresses corresponding to different interfaces of a node are resolved to one identifier); an anonymous node is represented as $*$ ("star") in the traces. For our formal analysis, we assign to each star in a trace set $\mathcal{T}$ a unique identifier $i$: $*_i$. (Note that except for the numbering of the stars, we allow identical copies of $T$ in $\mathcal{T}$, and we do not make any assumptions on the implications of identical traces: they may or may not describe the same paths.) Thus, a trace $T \in \mathcal{T}$ is a sequence of symbols taken from an alphabet $\Sigma = \mathcal{ID} \cup (\bigcup_i *_i)$, where $\mathcal{ID}$ is the set of non-anonymous node identifiers (IDs): $\Sigma$ is the union of the (anti-aliased) non-anonymous nodes and the set of all stars (with their unique identifiers) appearing in a trace set. The main challenge in topology inference is to determine which stars in the traces may originate from which anonymous nodes.

Henceforth, let $n = |\mathcal{ID}|$ denote the number of non-anonymous nodes and let $s = |\bigcup_i *_i|$ be the number of stars in $\mathcal{T}$; similarly, let $a$ denote the number of anonymous nodes in a topology. Let $N = n + s = |\Sigma|$ be the total number of symbols occurring in $\mathcal{T}$.

Table 1 gives some statistics for our variables from other studies [23, 15]: the number of traces $|\mathcal{T}|$ is between 372 and 27510, the number of named nodes $n$ is between 615 and 1077, the number of stars $s$ is between 62 and 4571, and the number of edges in the trace set varies between 1026 and 10011. (See Section 1.1 for a short summary on the studies collecting these trace sets.)

Clearly, the process of topology inference depends on the assumptions on the measurements. In the following, we postulate the fundamental axioms that guide the reconstruction. First, we make the assumption that each link of $G_0$ is visited by the measurement process, i.e., it appears as a transition in the trace set $\mathcal{T}$. In other words, we are only interested in inferring the (sub-)graph for which measurement data is available.

AXIOM 0 (*Complete Cover*): Each edge of $G_0$ appears at least once in some trace in $\mathcal{T}$.

The next fundamental axiom assumes that traces always represent paths on $G_0$.

AXIOM 1 (*Reality Sampling*): For every trace $T \in \mathcal{T}$, if the distance between two symbols $\sigma_1, \sigma_2 \in T$ is $d_T(\sigma_1, \sigma_2) = k$, then there exists a path (i.e., a walk without cycles) of length $k$ connecting two (named or anonymous) nodes $\sigma_1$ and $\sigma_2$ in $G_0$.

The following axiom captures the consistency of the routing protocol on which the traceroute probing relies. In the current Internet, policy routing is known to have in impact both on the route length [26] and on the convergence time [20].

AXIOM 2 ($\alpha$*-(Routing) Consistency*): There exists an $\alpha \in (0, 1]$ such that, for every trace $T \in \mathcal{T}$, if $d_T(\sigma_1, \sigma_2) = k$ for two entries $\sigma_1, \sigma_2$ in trace $T$, then the shortest path connecting the two (named or anonymous) nodes corresponding to $\sigma_1$ and $\sigma_2$ in $G_0$ has distance at least $\lceil \alpha k \rceil$.

Note that if $\alpha = 1$, the routing is a shortest path routing. Moreover, note that if $\alpha = 0$, there can be loops in the paths, and there are hardly any topological constraints, rendering almost any topology inferrable. (For example, the complete graph with one anonymous router is always a solution.)

Any topology $G$ which is consistent with these axioms (when applied to $\mathcal{T}$) is called *inferrable* from $\mathcal{T}$.

**Definition 1 (Inferrable Topologies)** A topology $G$ is ($\alpha$-consistently) *inferrable* from a trace set $\mathcal{T}$ if axioms AXIOM 0, AXIOM 1, and AXIOM 2 (with parameter $\alpha$) are fulfilled.

We will refer by $\mathcal{G}_\mathcal{T}$ to the set of topologies inferrable from $\mathcal{T}$. Please note the following important observation.

*Remark 1* In the absence of anonymous nodes, it holds that $G_0 \in \mathcal{G}_\mathcal{T}$, since $\mathcal{T}$ was generated from $G_0$ and AXIOM 0, AXIOM 1, and AXIOM 2 are fulfilled by definition. However, there are instances where an $\alpha$-consistent trace set for $G_0$ contradicts AXIOM 0: as trace needs to start and end with a named node, some edges cannot appear in an $\alpha$-consistent trace set $\mathcal{T}$. In the remainder of this article, we will only consider settings where $G_0 \in \mathcal{G}_\mathcal{T}$.

| data | # traces | n | s | # edges | # named edges | # src-dst pairs | avg trace length |
|---|---|---|---|---|---|---|---|
| routes1 [21] | 6219 | 746 | 513 | 2372 | 1576 | 893 | 14.81 |
| routes2 [21] | 27510 | 1077 | 4571 | 10011 | 2243 | 1417 | 15.57 |
| xprobes [21] | 6407 | 909 | 631 | 2688 | 1512 | 905 | 15.69 |
| xroutes.1 [21] | 615 | 673 | 62 | 1026 | 906 | 273 | 15.83 |
| PAM [14] | 372 | 1511 | 279 | 2685 | 2153 | 372 | 12.25 |

**Fig. 1** This table gives an overview of the order of magnitudes of our model parameters in real life data. The number of edges refers to the number of edges in the trace set (the underlying graphs are unknown). The number of named edges counts the number of edges between named nodes. Some trace sets contain multiple queries for the same source-destination pairs, whereas others consist of one trace per pair.

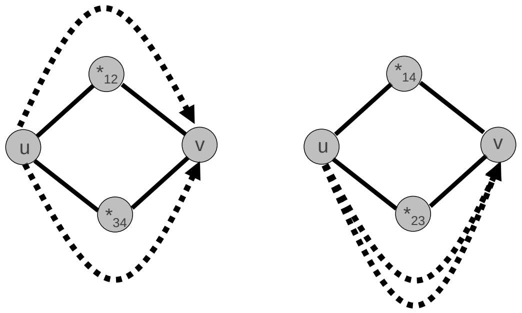


**Fig. 2** Two non-isomorphic inferred topologies, i.e., different mapping functions lead to these topologies.

The main objective of a topology inference algorithm ALG is to compute topologies which are consistent with these axioms. Concretely, ALG's input is the trace set $\mathcal{T}$ together with the parameter $\alpha$ specifying the assumed routing consistency. Essentially, the goal of any topology inference algorithm ALG is to compute a mapping of the symbols $\Sigma$ (appearing in $\mathcal{T}$) to nodes in an inferred topology $G$; or, in case the input parameters $\alpha$ and $\mathcal{T}$ are contradictory, reject the input. This mapping of symbols to nodes implicitly describes the edge set of $G$ as well: the edge set is unique as all the transitions of the traces in $\mathcal{T}$ are now unambiguously tied to two nodes.

So far, we have ignored an important and non-trivial question: When are two topologies $G_1, G_2 \in \mathcal{G}_{\mathcal{T}}$ different (and hence appear as two independent topologies in $\mathcal{G}_{\mathcal{T}}$)? In this article, we pursue the following approach: We are not interested in purely topological isomorphisms, but we care about the identifiers of the non-anonymous nodes, i.e., we are interested in the locations of the non-anonymous nodes and their distance to other nodes. For anonymous nodes, the situation is slightly more complicated: one might think that as the nodes are anonymous, their "names" do not matter. Consider however the example in Figure 2: the two inferrable topologies have two anonymous nodes, one where $\{*_1, *_2\}$ plus $\{*_3, *_4\}$ are merged into one node each in the inferrable topology and one where $\{*_1, *_4\}$ plus $\{*_2, *_3\}$ are merged into one node each in the inferrable topology. In this article, we regard the two topologies as different, for the following reason: Assume that there are two paths in the network, one $u \rightsquigarrow *_2 \rightsquigarrow v$ (e.g., during day time) and one $u \rightsquigarrow *_3 \rightsquigarrow v$ (e.g., at night); clearly, this traffic has different consequences and hence we want to be able to distinguish between the two topologies described above. In other words, our notion of isomorphism of inferred topologies is *path-aware*.

It is convenient to introduce the following MAP function. Essentially, an inference algorithm computes such a mapping.

**Definition 2 (Mapping Function MAP)** Let $G = (V, E) \in \mathcal{G}_{\mathcal{T}}$ be a topology inferrable from $\mathcal{T}$. A topology inference algorithm describes a surjective mapping function MAP : $\Sigma \rightarrow V$. For the set of non-anonymous nodes in $\Sigma$, the mapping function is bijective; and each star is mapped to exactly one node in $V$, but multiple stars may be assigned to the same node. Note that for any $\sigma \in \Sigma$, MAP($\sigma$) uniquely identifies a node $v \in V$. More specifically, we assume that MAP assigns labels to the nodes in $V$: in case of a named node, the label is simply the node's identifier; in case of anonymous nodes, the label is $*_\beta$, where $\beta$ is the concatenation of the *sorted* indices of the stars which are merged into node $*_\beta$.

With this definition, two topologies $G_1, G_2 \in \mathcal{G}_{\mathcal{T}}$ differ if and only if they do not describe the identical (MAP-) labeled topology. We will use this MAP function also for $G_0$, i.e., we will write MAP($\sigma$) to refer to a symbol $\sigma$'s corresponding node in $G_0$.

AXIOM 1 implies a natural way to merge traces to derive additional bounds on path lengths.

**Lemma 1** *For two traces $T_1, T_2 \in \mathcal{T}$ for which $\exists \sigma_1, \sigma_2, \sigma_3$, where $\sigma_2$ refers to a named node, such that $d_{T_1}(\sigma_1, \sigma_2) = i$ and $d_{T_2}(\sigma_2, \sigma_3) = j$, it holds that the distance between two nodes $u$ and $v$ corresponding to $\sigma_1$ and $\sigma_2$, respectively, in $G_0$, is at most $d_{G_0}(\sigma_1, \sigma_3) \leq i + j$.*

*Proof* Let $\mathcal{T}$ be a trace set, and $G \in \mathcal{G}_{\mathcal{T}}$. Let $\sigma_1, \sigma_2, \sigma_3$ s.t. $\exists T_1, T_2 \in \mathcal{T}$ with $\sigma_1 \in T_1, \sigma_3 \in T_2$ and $\sigma_2 \in T_1 \cap T_2$. Let $i = d_{T_1}(\sigma_1, \sigma_2)$ and $j = d_{T_2}(\sigma_1, \sigma_3)$. Since any inferrable topology $G$ fulfills AXIOM 1, there is a path $\pi_1$ of length at most $i$ between the nodes corresponding to $\sigma_1$ and $\sigma_2$ in $G$ and a path $\pi_2$ of length at most $j$ between the nodes corresponding to $\sigma_2$ and $\sigma_3$ in $G$. The combined path can only be shorter, and hence the claim follows. $\square$

In the remainder of this article, we will often assume that AXIOM 0 is given. Thus, to prove that a topology is inferrable from a trace set it is sufficient to show that AXIOM 1 and AXIOM 2 are satisfied.

## 3 Inferrable Topologies

What insights can be obtained from topology inference with minimal assumptions, i.e., with our axioms? Or what is the structure of the inferrable topology set $\mathcal{G}_\mathcal{T}$? We first make some general observations and then examine different graph metrics in more detail.

### 3.1 Basic Observations

Although the generation of the entire topology set $\mathcal{G}_\mathcal{T}$ may be computationally hard, some instances of $\mathcal{G}_\mathcal{T}$ can be computed efficiently. The simplest possible inferrable topology is the so-called *canonic graph* $G_C$: the topology which assumes that all stars in the traces refer to different anonymous nodes. In other words, if a trace set $\mathcal{T}$ contains $n = |\mathcal{ID}|$ named nodes and $s$ stars, $G_C$ will contain $|V(G_C)| = N = n + s$ nodes.

**Definition 3 (Canonic Graph $G_C$)** The *canonic graph* is defined by $G_C(V_C, E_C)$ where $V_C = \Sigma$ is the set of (anti-aliased) nodes appearing in $\mathcal{T}$ (where each star is considered a unique anonymous node) and where $\{\sigma_1, \sigma_2\} \in E_C \Leftrightarrow \exists T \in \mathcal{T}, T = (\ldots, \sigma_1, \sigma_2, \ldots)$, i.e., $\sigma_1$ follows after $\sigma_2$ in some trace $T$ ($\sigma_1, \sigma_2 \in T$ can be either non-anonymous nodes or stars). Let $d_C(\sigma_1, \sigma_2)$ denote the *canonic distance* between two nodes, i.e., the length of a shortest path in $G_C$ between the nodes $\sigma_1$ and $\sigma_2$.

Note that $G_C$ is indeed one inferrable topology. In this case, MAP : $\Sigma \to \Sigma$ is the identity function.

**Theorem 1** *$G_C$ is inferrable from $\mathcal{T}$.*

*Proof* Fix $\mathcal{T}$. We have to prove that $G_C$ fulfills AXIOM 0, AXIOM 1 and AXIOM 2.

AXIOM 0: The axiom holds trivially: only edges from the traces are used in $G_C$.

AXIOM 1: Let $T \in \mathcal{T}$ and $\sigma_1, \sigma_2 \in T$. Let $k = d_T(\sigma_1, \sigma_2)$. We show that $G_C$ fulfills AXIOM 1, namely, there exists a path of length $k$ in $G_C$. Induction on $k$: ($k = 1$:) By the definition of $G_C$, $\{\sigma_1, \sigma_2\} \in E_C$ thus there exists a path of length one between $\sigma_1$ and $\sigma_2$. ($k > 1$:) Suppose AXIOM 1 holds up to $k-1$. Let $\sigma'_1, \ldots, \sigma'_{k-1}$ be the intermediary nodes between $\sigma_1$ and $\sigma_2$ in $T$: $T = (\ldots, \sigma_1, \sigma'_1, \ldots, \sigma'_{k-1}, \sigma_2, \ldots)$. By the induction hypothesis, in $G_C$ there is a path of length $k-1$ between $\sigma_1$ and $\sigma'_{k-1}$. Let $\pi$ be this path. By definition of $G_C$, $\{\sigma'_{k-1}, \sigma_2\} \in E_C$. Thus appending $(\sigma'_{k-1}, \sigma_2)$ to $\pi$ yields the desired path of length $k$ linking $\sigma_1$ and $\sigma_2$: AXIOM 1 thus holds up to $k$.

AXIOM 2: We have to show that $d_T(\sigma_1, \sigma_2) = k \Rightarrow d_C(\sigma_1, \sigma_2) \geq \lceil \alpha \cdot k \rceil$. By contradiction, suppose that $G_C$ does not fulfill AXIOM 2 with respect to $\alpha$. So there exists $k' < \lceil \alpha \cdot k \rceil$ and $\sigma_1, \sigma_2 \in V_C$ such that $d_C(\sigma_1, \sigma_2) = k'$. Let $\pi$ be a shortest path between $\sigma_1$ and $\sigma_2$ in $G_C$. Let $(T_1, \ldots, T_\ell)$ be the corresponding (maybe repeating) traces covering this path $\pi$ in $G_C$. Let $T_i \in (T_1, \ldots, T_\ell)$, and let $s_i$ and $e_i$ be the corresponding start and end nodes of $\pi$ in $T_i$. We will show that this path $\pi$ implies the existence of a path in $G_0$ which violates $\alpha$-consistency. Since $G_0$ is inferrable, $G_0$ fulfills AXIOM 2, thus we have: $d_C(\sigma_1, \sigma_2) = \sum_{i=1}^{\ell} d_{T_i}(s_i, e_i) = k' < \lceil \alpha \cdot k \rceil \leq d_{G_0}(\sigma_1, \sigma_2)$ since $G_0$ is $\alpha$-consistent. However, $G_0$ also fulfills AXIOM 1, thus $d_{T_i}(s_i, e_i) \geq d_{G_0}(s_i, e_i)$. Thus $\sum_{i=1}^{\ell} d_{G_0}(s_i, e_i) \leq \sum_{i=1}^{\ell} d_{T_i}(s_i, e_i) < d_{G_0}(\sigma_1, \sigma_2)$: we have constructed a path from $\sigma_1$ to $\sigma_2$ in $G_0$ whose length is shorter than the distance between $\sigma_1$ and $\sigma_2$ in $G_0$, leading to the desired contradiction. $\square$

Theorem 1 implies that with our axioms the canonic graph is one of the possible topologies that could lead to a given trace set. However, it does not imply that the stars of traces from a given topology represent different nodes.

$G_C$ can be computed efficiently from $\mathcal{T}$: represent each non-anonymous node and star as a separate node, and for any pair of consecutive entries (i.e., nodes) in a trace, add the corresponding link. The time complexity of this construction is linear in the size of $\mathcal{T}$. Also note that there is no inferrable graph in $\mathcal{G}_\mathcal{T}$ having a larger diameter than $G_C$.

With the definition of the canonic graph, we can derive the following lemma which establishes a necessary condition when two stars cannot represent the same node in $G_0$ from constraints on the routing paths. This is useful for the characterization of inferred topologies.

**Lemma 2** *Let $*_1, *_2$ be two stars occurring in some traces in $\mathcal{T}$. $*_1, *_2$ cannot be mapped to the same node, i.e.,* MAP$(*_1) \neq$ MAP$(*_2)$*, without violating the axioms in the following conflict situations:*

*(i) if $*_1 \in T_1$ and $*_2 \in T_2$, and $T_1$ describes too a long path between anonymous node* MAP$(*_1)$ *and non-anonymous node $u$, i.e., $\lceil \alpha \cdot d_{T_1}(*_1, u) \rceil > d_C(u, *_2)$.*

*(ii) if $*_1 \in T_1$ and $*_2 \in T_2$, and there exists a trace $T$ that contains a path between two non-anonymous nodes $u$ and $v$ and $\lceil \alpha \cdot d_T(u, v) \rceil > d_C(u, *_1) + d_C(v, *_2)$.*

*Proof* The first proof is by contradiction. Assume MAP$(*_1)$ = MAP$(*_2)$ represents the same node $v$ of $G_0$, and that $\lceil \alpha \cdot d_{T_1}(v, u) \rceil > d_C(u, v)$. Then we know from AXIOM 2 that $d_C(v, u) \geq d_{G_0}(v, u) \geq \lceil \alpha \cdot d_{T_1}(u, v) \rceil > d_C(v, u)$, which yields the desired contradiction.

Similarly for the second proof, assume for the sake of contradiction that $\textsc{Map}(*_1) = \textsc{Map}(*_2)$ represents the same node $w$ of $G_0$, and that $\lceil \alpha \cdot d_T(u,v) \rceil > d_C(u,*_1) + d_C(v,*_2) \geq d_{G_0}(u,w) + d_{G_0}(v,w)$. Due to the triangle inequality, we have that $d_{G_0}(u,w) + d_{G_0}(v,w) \geq d_{G_0}(u,v)$ and hence, $\lceil \alpha \cdot d_T(u,v) \rceil > d_{G_0}(u,v)$, which contradicts the fact that $G_0$ is inferrable (Remark 1). □

Lemma 2 can be applied to show that a topology is not inferrable from a given trace set because it merges (i.e., maps to the same node) two stars in a manner that violates the axioms. Let us introduce a useful concept for our analysis: the *star graph* that describes the conflicts between stars.

**Definition 4 (Star Graph $G_*$)** The *star graph* $G_*(V_*, E_*)$ consists of vertices $V_*$ representing stars in traces, i.e., $V_* = \bigcup_i *_i$. Two vertices are connected if and only if they must differ according to Lemma 2, i.e., $\{*_1, *_2\} \in E_*$ if and only if at least one of the conditions of Lemma 2 hold for $*_1, *_2$.

Note that the star graph $G_*$ is unique and can be computed efficiently for a given trace set $\mathcal{T}$: Conditions (i) and (ii) can be checked by computing $G_C$. However, note that while $G_*$ specifies some stars which cannot be merged, the construction is not sufficient: as Lemma 2 is based on $G_C$, additional links might be needed to characterize the set of inferrable and $\alpha$-consistent topologies $\mathcal{G}_\mathcal{T}$ exactly. In other words, a topology $G$ obtained by merging stars that are adjacent in $G_*$ is never inferrable ($G \notin \mathcal{G}_\mathcal{T}$); however, merging non-adjacent stars does not guarantee that the resulting topology is inferrable.

What do star graphs look like? The answer is *arbitrarily*: the following lemma states that the set of possible star graphs is equivalent to the class of general graphs. This claim holds for any $\alpha$.

**Lemma 3** *For any graph $G = (V,E)$, there exists a trace set $\mathcal{T}$ such that $G$ is the star graph for $\mathcal{T}$.*

*Proof* First we show how to construct a topology $G_0 = (V_0, E_0)$ for a given star graph $G = (V,E)$ and then describe a trace set of $G_0$ that generates the star graph $G$. The node set $V_0$ consists of $|V|$ anonymous nodes and $|V| \cdot (1+\tau)$ named nodes, where $\tau = \lceil 3/(2\alpha) - 1/2 \rceil$. The first building block of $G_0$ is a copy of $G$. To each node $v_i$ in the copy of $G$ we add a chain consisting of $2+\tau$ nodes, first appending $\tau$ non-anonymous nodes $w_{(i,k)}$ where $1 \leq k \leq \tau$, followed by an anonymous node $u_i$ and finally a named node $w_{(i,\tau+1)}$. More formally, we can describe the link set as $E_0 = E \cup \bigcup_{i=1}^{|V|} \big(\{v_i, w_{(i,1)}\}, \{w_{(i,1)}, w_{(i,2)}\}, \ldots, \{w_{(i,\tau)}, u_i\}, \{u_i, w_{(i,\tau+1)}\}\big)$. The trace set $\mathcal{T}$ consists of the following $|V| + |E|$ shortest path traces: the traces $T_\ell$ for $\ell \in \{1, \ldots, |V|\}$, are given by $T_\ell(w_{(\ell,\tau)}, w_{(\ell,\tau+1)})$ (for each node in $V$), and the traces $T_\ell$ for $\ell \in \{|V|+1, \ldots, |V|+|E|\}$, are given by $T_\ell(w_{(i,\tau)}, w_{(j,\tau)})$ for each link $\{v_i, v_j\}$ in $E$. Note that $G_0 = G_C$ as each star appears as a separate anonymous node. The star graph $G_*$ corresponding to this trace set contains the $|V|$ nodes $*_i$ (corresponding to $u_i$). In order to prove the claim of the lemma we have to show that two nodes $*_i, *_j$ are conflicting according to Lemma 2 if and only if there is a link $\{v_i, v_j\}$ in $E$. Case $(i)$ does not apply because the minimum distance between any two nodes in the canonic graph is at least one, and $\lceil \alpha \cdot d_{T_i}(*_i, w_{(i,\tau)}) \rceil = 1$ and $\lceil \alpha \cdot d_{T_i}(*_i, w_{(i,\tau+1)}) \rceil = 1$. It remains to examine Case $(ii)$: "⇒" if $\textsc{Map}(*_i) = \textsc{Map}(*_j)$ there would be a path of length two between $w_{(i,\tau)}$ and $w_{(j,\tau)}$ in the topology generated by $\textsc{Map}$; the trace set however contains a trace $T_\ell(w_{(i,\tau)}, w_{(j,\tau)})$ of length $2\tau + 1$. So $\lceil \alpha \cdot d_{T_\ell}(w_{(i,\tau)}, w_{(j,\tau)}) \rceil = \lceil \alpha \cdot (2\tau+1) \rceil = \lceil \alpha \cdot (2\lceil 3/(2\alpha) - 1/2 \rceil + 1 \rceil) \geq 3$, which violates the $\alpha$-consistency (Lemma 2 (ii)) and hence $\{*_i, *_j\} \in E_*$ and $\{v_i, v_j\} \in E$. "⇐": if $\{v_i, v_j\} \notin E$, there is no trace $T(w_{(i,\tau)}, w_{(j,\tau)})$, thus we have to prove that no trace $T_\ell(w_{(i',\tau)}, w_{(j',\tau)})$ with $i' \neq i$ and $j' \neq j$ and $j' \neq i$ leads to a conflict between $*_i$ and $*_j$. We show that an even more general statement is true, namely that for any pair of distinct non-anonymous nodes $x_1, x_2$, where $x_1, x_2 \in \{v_{i'}, v_{j'}, w_{(i',k)}, w_{(j',k)} | 1 \leq k \leq \tau+1, i' \neq i, j' \neq i, j' \neq j\}$, it holds that $\lceil \alpha \cdot d_C(x_1, x_2) \rceil \leq d_C(x_1, *_i) + d_C(x_2, *_j)$. Since $G_C = G_0$ and the traces contain shortest paths only, the trace distance between two nodes in the same trace is the same as the distance in $G_C$. The following tables contain the relevant lower bounds on distances in $G_C$ and $\mu(x_1, x_2) = d_C(x_1, *_i) + d_C(x_2, *_j)$.

If $x_1 \neq w_{(j',k_2)}$ then it holds for all $x_1, x_2$ that $d_{T_\ell}(x_1, x_2) \leq 2\tau + 1$ whereas $\mu(x_1, x_2) = d_C(x_1, *_i) + d_C(x_2, *_j) \geq 2\tau + 2$. In all other cases it holds at least that $d_C(x_1, x_2) < \mu(x_1, x_2)$. Thus $\lceil \alpha \cdot d_C(x_1, x_2) \rceil \leq d_C(x_1, *_i) + d_C(x_2, *_j)$. Consequently, we have conflicts if and only if $\{v_i, v_j\} \in E$, which concludes the proof. □

The problem of computing inferrable topologies is related to the vertex colorings of the star graphs. We will use the following definition which relates a vertex coloring of $G_*$ to an inferrable topology $G$ by contracting independent stars in $G_*$ to become one anonymous node in $G$. For example, observe that a maximum coloring treating every star in the trace as a separate anonymous node describes the inferrable topology $G_C$.

**Definition 5 (Coloring-Induced Graph)** Let $\gamma$ denote a coloring of $G_*$ which assigns colors $1, \ldots, k$ to the vertices of $G_*$: $\gamma : V_* \to \{1, \ldots, k\}$. We require that $\gamma$ is a proper coloring of $G_*$, i.e., that different anonymous nodes are assigned different colors: $\{u, v\} \in E_* \Rightarrow \gamma(u) \neq \gamma(v)$. $G_\gamma$ is defined as the topology *induced* by $\gamma$. $G_\gamma$ describes the graph $G_C$ where nodes of the same color are contracted:

| $d_C(\cdot,\cdot) \geq$ | $v_{i'}$ | $v_{j'}$ | $w_{(i',k_1)}$ | $w_{(j',k_1)}$ |
|---|---|---|---|---|
| $v_{i'}$ | $0$ | $1$ | $k_1$ | $k_1+1$ |
| $v_{j'}$ | $1$ | $0$ | $k_1+1$ | $k_1$ |
| $w_{(i',k_2)}$ | $k_2$ | $k_2+1$ | $\|k_2-k_1\|$ | $k_1+1+k_2$ |
| $w_{(j',k_2)}$ | $k_2+1$ | $k_2$ | $k_1+1+k_2$ | $\|k_2-k_1\|$ |
| $*_i$ | $\tau+2$ | $\tau+1$ | $2+\tau+k_1$ | $\tau-k_1+1$ |
| $*_j$ | $\tau+2$ | $\tau+2$ | $2+\tau+k_1$ | $2+\tau+k_1$ |

| $\mu(\cdot,\cdot) \geq$ | $v_{i'}$ | $v_{j'}$ | $w_{(i',k_1)}$ | $w_{(j',k_1)}$ |
|---|---|---|---|---|
| $v_{i'}$ | $2\tau+4$ | $2\tau+3$ | $4+2\tau+k_1$ | $4+2\tau+k_1$ |
| $v_{j'}$ | $2\tau+3$ | $2\tau+4$ | $2\tau+3+k_1$ | $3+2\tau+k_1$ |
| $w_{(i',k_2)}$ | $4+2\tau+k_2$ | $4+2\tau+k_2$ | $4+2\tau+k_1+k_2$ | $4+2\tau+k_1+k_2$ |
| $w_{(j',k_2)}$ | $2\tau-k_2+3$ | $2\tau-k_2+3$ | $2\tau+3+k_1-k_2$ | $2\tau+k_1-k_2+3$ |

**Fig. 3** Proof of Lemma 3: lower bounds for the distances in $G_C$, and lower bounds for $\mu(x_1,x_2) = d_C(x_1,*_i) + d_C(x_2,*_j)$.

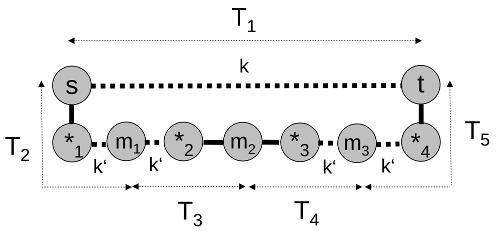


**Fig. 4 Visualization for proof of Lemma 4. Solid lines denote links, dashed lines denote paths (of annotated length).**

two vertices $u$ and $v$ represent the same node in $G_\gamma$, i.e., $\text{MAP}(*_i) = \text{MAP}(*_j)$, if and only if $\gamma(*_i) = \gamma(*_j)$.

The following two lemmas establish an intriguing relationship between colorings of $G_*$ and inferrable topologies. Also note that Definition 5 implies that two different colorings of $G_*$ define two non-isomorphic inferrable topologies.

We first show that while a coloring-induced topology always fulfills AXIOM 1, the routing consistency is sacrificed.

**Lemma 4** *Let $\gamma$ be a proper coloring of $G_*$. The coloring induced topology $G_\gamma$ is a topology fulfilling* AXIOM 2 *with a routing consistency of $\alpha' > 0$, for an arbitrarily small $\alpha'$.*

*Proof* We have to show that the paths in the traces correspond to paths in $G_\gamma$. Let $T \in \mathcal{T}$, and $\sigma_1, \sigma_2 \in T$. Let $\pi$ be the sequence of nodes in $T$ connecting $\sigma_1$ and $\sigma_2$. This is also a path in $G_\gamma$: since $\alpha > 0$, for any two symbols $\sigma_1, \sigma_2 \in T$, it holds that $\text{MAP}(\sigma_1) \neq \text{MAP}(\sigma_2)$ as $\alpha > 0$. Since the traces we consider are finite, the ratio of the shortest path between two nodes and length of the longest trace always exceeds zero and thus $\alpha' > 0$.

We now construct an example showing that the $\alpha'$ for which $G_\gamma$ fulfills AXIOM 2 can be arbitrarily small. Consider the graph represented in Figure 4. Let $T_1 = (s,\ldots,t), T_2 = (s,*_1,\ldots,m_1), T_3 = (m_1,\ldots,*_2,m_2), T_4 = (m_2,*_3,\ldots,m_3), T_5 = (m_3,\ldots,*_4,t)$. We assume $\alpha = 1$. By changing parameters $k = d_C(s,t)$ and $k' = d_C(m_1,*_1) = d_C(m_1,*_2) = d_C(m_3,*_3) = d_C(m_3,*_4)$, we can modulate the links of the corresponding star graph $G_*$. Using $d_{T_1}(s,t) = k$, observe that $k > 2 \Leftrightarrow \{*_1,*_4\} \in E_*$. Similarly, $k > 2(k'+1) \Leftrightarrow \{*_1,*_3\} \in E_* \wedge \{*_2,*_4\} \in E_*$ and $k > 2(k'+2) \Leftrightarrow \{*_1,*_2\} \in E_* \wedge \{*_3,*_4\} \in E_*$. Taking $k = 2k'+4$, we thus have $E_* = \{\{*_1,*_3\},\{*_2,*_4\},\{*_1,*_4\}\}$.

Thus, we here construct a situation where $*_1$ and $*_2$ as well as $*_3$ and $*_4$ can be merged without breaking the consistency requirement, but where merging both simultaneously leads to a topology $G'$ that is only $4/k$-consistent, since $d_{G'}(s,t) = 4$. This ratio can be made arbitrarily small provided we choose $k' = (k-4)/2$. $\square$

An inferrable topology always defines a proper coloring on $G_*$.

**Lemma 5** *Let $\mathcal{T}$ be a trace set and $G_*$ its corresponding star graph. If a topology $G$ is inferrable from $\mathcal{T}$, then $G$ induces a proper coloring on $G_*$.*

*Proof* For any $\alpha$-consistent inferrable topology $G$ there exists some mapping function MAP that assigns each symbol of $\mathcal{T}$ to a corresponding node in $G$ (cf Definition 2), and this mapping function gives a coloring on $G_*$ (i.e., merged stars appear as nodes of the same color in $G_*$). The coloring must be proper: due to Lemma 2, an inferrable topology can never merge adjacent nodes of $G_*$. $\square$

The colorings of $G_*$ allow us to derive an upper bound on the cardinality of $\mathcal{G_T}$.

**Theorem 2** *Given a trace set $\mathcal{T}$ sampled from a network $G_0$ and $\mathcal{G_T}$, the set of topologies inferrable from $\mathcal{T}$, it holds that:*

$$\sum_{k=\gamma(G_*)}^{|V_*|} P(G_*,k)/k! \geq |\mathcal{G_T}|,$$

*where $\gamma(G_*)$ is the chromatic number of $G_*$ and $P(G_*, k)$ is the number of colorings of $G_*$ with k colors (known as the* chromatic polynomial *of $G_*$).*

*Proof* The proof follows directly from Lemma 5 which shows that each inferred topology has proper colorings, and the fact that a coloring of $G_*$ cannot result in two different inferred topologies, as the coloring uniquely describes which stars to merge (Lemma 4). In order to account for isomorphic colorings, we need to divide by the number of color permutations. $\square$

Note that the fact that $G_*$ can be an arbitrary graph (Lemma 3) implies that we cannot exploit some special properties of $G_*$ to compute colorings of $G_*$ and $\gamma(G_*)$. Also note that the exact computation of the upper bound is hard, since the minimal coloring as well as the chromatic polynomial of $G_*$ (in P♯) is needed. To complement the upper bound, we note that star graphs with a small number of conflict edges can indeed result in a large number of inferred topologies.

**Theorem 3** *Regardless of $\alpha > 0$, there is a trace set for which the number of non-isomorphic colorings of $G_*$ equals $|\mathcal{G}_\mathcal{T}| \leq B_s$, where $\mathcal{G}_\mathcal{T}$ is the set of inferrable and $\alpha$-consistent topologies, $s$ is the number of stars in $\mathcal{T}$, and $B_s$ is the* Bell number *of $s$. Such a trace set can originate from a $G_0$ network with one anonymous node only.*

*Proof* Consider a trace set $\mathcal{T} = \{(\sigma_i, *_i, \sigma'_i)_{i=1,\ldots,s}\}$ (e.g., obtained from exploring a topology $G_0$ where one anonymous center node is connected to $2s$ named nodes). The trace set does not impose any constraints on how the stars relate to each other, and hence, $G_*$ does not contain any edges at all; even when stars are merged, there are no constraints on how the stars relate to each other. Therefore, the star graph for $\mathcal{T}$ has $B_s = \sum_{j=0}^{s} S_{(s,j)}$ colorings, where $S_{(s,j)} = 1/j! \cdot \sum_{\ell=0}^{j} (-1)^\ell \binom{j}{\ell} (j-\ell)^s$ is the number of ways to group $s$ nodes into $j$ different, disjoint non-empty subsets (known as the *Stirling number of the second kind*). Each of these colorings also describes a distinct inferrable topology as MAP assigns unique labels to anonymous nodes stemming from merging a group of stars (cf Definition 2). $\square$

Using these observations it is now possible to design an algorithm extracting $\mathcal{G}_\mathcal{T}$: first, $i$) construct $G_c$ and $G_*$ using $\mathcal{T}$ and the parameter $\alpha$. Then, $ii$) compute all the non isomorphic proper colorings of $G_*$. Each such coloring $\gamma$ defines which vertices to merge in $G_c$ to obtain a color-induced topology $G_\gamma$. Finally, $iii$) for each color-induced graph $G_\gamma$, test whether it is $\alpha$-consistent with respect to $\mathcal{T}$. Note that each color of a proper coloring $\gamma$ yields an anonymous router in $G_\gamma$. Thus, if one is only interested in minimal topologies, it is possible to compute only the minimal colorings of $G_*$ on step $ii$). More formally, Algorithm 1 summarizes the steps required to produce the set of all inferrable topologies $\mathcal{G}_\mathcal{T}$.

**Algorithm 1** Given traces $\mathcal{T}$:

```
Compute G_* and G_C
𝒢_𝒯 ← ∅
for all proper colorings γ of G_* do
    G_γ ← G_C
    for all pairs {*_i, *_j} do
        if γ(*_i) = γ(*_i) then
            merge *_i and *_j in G_γ
        end if
    end for
    𝒢_𝒯 ← 𝒢_𝒯 ∪ G_γ
end for
return 𝒢_𝒯
```

## 3.2 Properties

Even if the number of inferrable topologies is large, studying trace sets can still be useful if one is mainly interested in some properties of $G_0$ and if the ensemble $\mathcal{G}_\mathcal{T}$ is homogenous with respect to these properties; for example, if "most" of the instances in $\mathcal{G}_\mathcal{T}$ the properties are close to $G_0$, it may be an option to conduct an efficient sampling analysis on random representatives. Therefore, in the following, we will take a closer look on how much the members of $\mathcal{G}_\mathcal{T}$ differ in various aspects.

Important metrics to characterize inferrable topologies are, for instance, the graph size, the diameter $\textsc{Diam}(\cdot)$, the number of triangles $C_3(\cdot)$ of $G$. In the following, let $G_1 = (V_1, E_1), G_2 = (V_2, E_2) \in \mathcal{G}_\mathcal{T}$ be two arbitrary representatives of $\mathcal{G}_\mathcal{T}$.

The possible difference and ratio of inferrable graph sizes is at most linear in the number of stars.

**Theorem 4** *It holds that $|V_1| - |V_2| \leq s - \gamma(G_*) \leq s - 1$ and $|V_1|/|V_2| \leq (n+s)/(n+\gamma(G_*)) \leq (2+s)/3$. Moreover, $|E_1| - |E_2| \leq 2(s - \gamma(G_*))$ and $|E_1|/|E_2| \leq (\nu + 2s)/(\nu + 2) \leq s$, where $\nu$ denotes the number of edges between non-anonymous nodes. There are traces with inferrable topology $G_1, G_2$ reaching these bounds.*

*Proof* In the worst-case, each star in the trace represents a different node in $G_1$, so the maximal number of nodes in any topology in $\mathcal{G}_\mathcal{T}$ is the total number of non-anonymous nodes plus the total number of stars in $\mathcal{T}$. This number of nodes is reached in the topology $G_C$. According to Definition 4, only non-adjacent stars in $G_*$ can represent the same node in an inferrable topology. Thus, the stars in trace $\mathcal{T}$ must originate from at least $\gamma(G_*)$ different nodes. As a consequence $|V_1| - |V_2| \leq s - \gamma(G_*)$, which can reach $s - 1$ for a trace set $\mathcal{T} =$

$\{T_i = (v, *_i, w) | 1 \le i \le s\}$. Analogously, $|V_1|/|V_2| \le (n+s)/(n+\gamma(G_*)) \le (2+s)/3$.

Observe that each occurrence of a node in a trace describes at most two edges. If all anonymous nodes are merged into $\gamma(G_*)$ nodes in $G_1$ and are separate nodes in $G_2$ the difference in the number of edges is at most $2(s-\gamma(G_*))$. Analogously, $|E_1|/|E_2| \le (\nu+2s)/(\nu+2) \le s$. The trace set $\mathcal{T} = \{T_i = (v, *_i, w) | 1 \le i \le s\}$ reaches this bound. □

Observe that inferrable topologies can also differ in the number of connected components. This implies that the shortest distance between two named nodes can differ arbitrarily between two representatives in $\mathcal{G}_\mathcal{T}$.

**Theorem 5** *Let* $\text{COMP}(G)$ *denote the number of connected components of a topology G. Then,* $|\text{COMP}(G_1) - \text{COMP}(G_2)| \le n/2$. *There are traces with inferrable topology* $G_1, G_2$ *reaching these bounds.*

*Proof* Consider the trace set $\mathcal{T} = \{T_i, i = 1 \ldots \lfloor n/2 \rfloor\}$ in which $T_i = \{n_{2i}, *_i, n_{2i+1}\}$. Since $i \ne j \Rightarrow T_i \cap T_j = \emptyset$, we have $|E_*| = 0$. Take $G_1$ as the 1-coloring of $G_*$: $G_1$ is a topology with one anonymous node connected to all named nodes. Take $G_2$ as the $\lfloor n/2 \rfloor$-coloring of the star graph: $G_2$ has $\lfloor n/2 \rfloor$ distinct connected components (consisting of three nodes).

*Upper bound:* For the sake of contradiction, suppose $\exists \mathcal{T}$ s.t. $|\text{COMP}(G_1) - \text{COMP}(G_2)| > \lfloor n/2 \rfloor$. Let us assume that $G_1$ has the most connected components: $G_1$ has at least $\lfloor n/2 \rfloor + 1$ more connected components than $G_2$. Let $C$ refer to a connected component of $G_2$ whose nodes are not connected in $G_1$. This means that $C$ contains at least one anonymous node. Thus, $C$ contains at least two named nodes (since a trace $T$ cannot start or end by a star). There must exist at least $\lfloor n/2 \rfloor + 1$ such connected component $C$. Thus $G_2$ has to contain at least $2(\lfloor n/2 \rfloor + 1) \ge n+1$ named nodes. Contradiction. □

Another aspect of the usefulness of topology inference depends on the distortion of shortest paths.

**Definition 6 (Stretch)** The maximal ratio of the distance of two non-anonymous nodes in $G_0$ and a connected topology $G$ is called the *stretch* $\rho$:

$$\rho = \max_{u,v \in \mathcal{ID}(G_0)} \max \left\{ \frac{d_{G_0}(u,v)}{d_G(u,v)}, \frac{d_G(u,v)}{d_{G_0}(u,v)} \right\}.$$

From Theorem 5 we already know that inferrable topologies can differ in the number of connected components, and hence, the distance and the stretch between nodes can be arbitrarily wrong. Hence, in the following, we will focus on connected graphs only. However, even if two nodes are connected, their distance in inferrable topologies can be much longer or shorter than in $G_0$.

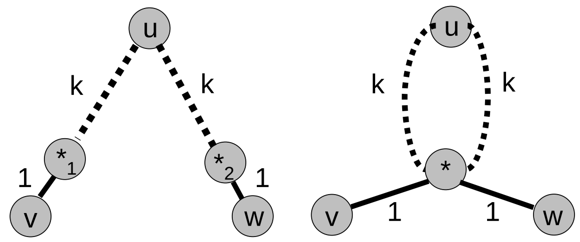


**Fig. 5** Due to the lack of a trace between $v$ and $w$, the stretch of an inferred topology can be large.

Figure 5 gives an example. Both topologies are inferrable from the traces $T_1 = (v, *, v_1, \ldots, v_k, u)$ and $T_2 = (w, *, w_1, \ldots, w_k, u)$. One inferrable topology is the canonic graph $G_C$ (Figure 5 *left*), whereas the other topology merges the two anonymous nodes (Figure 5 *right*). The distances between $v$ and $w$ are $2(k+2)$ and 2, respectively, implying a stretch of $k+2$.

**Theorem 6** *Let* $u$ *and* $v$ *be two arbitrary named nodes in the connected topologies* $G_1$ *and* $G_2$. *Then, even for only two stars in the trace set, it holds for the stretch that* $\rho \le (N-1)/2$. *There are traces with inferrable topologies* $G_1, G_2$ *reaching these bounds.*

*Proof* A "lower bound" example follows from Figure 5. Essentially, this is also the worst case: note that the difference in the shortest distance between a pair of nodes $u$ and $v$ in $G_1$ and $G_2$ is only greater than 0 if the shortest path between them involves at least one anonymous node. Hence the shortest distance between such a pair is two. The longest shortest distance between the same pair of nodes in another inferred topology visits all nodes in the network, i.e., its length is bounded by $N-1$. □

We now turn our attention to the diameter and the degree.

**Theorem 7** *For connected topologies* $G_1, G_2$ *it holds that* $\text{DIAM}(G_1) - \text{DIAM}(G_2) \le (s-1)/s \cdot \text{DIAM}(G_C) \le (s-1)(N-1)/s$ *and* $\text{DIAM}(G_1)/\text{DIAM}(G_2) \le s$, *where* DIAM *denotes the graph diameter and* $\text{DIAM}(G_1) > \text{DIAM}(G_2)$. *There are instances* $G_1, G_2$ *that reach these bounds.*

*Proof Upper bound:* As $G_C$ does not merge any stars, it describes the network with the largest diameter. Let $\pi$ be a longest path between two nodes $u$ and $v$ in $G_C$. In the extreme case, $\pi$ is the only path determining the network diameter and $\pi$ contains all star nodes. Then, the graph where all $s$ stars are merged into one anonymous node has a minimal diameter of at least $\text{DIAM}(G_C)/s$.

*Instances meeting the bound:* Consider the trace set $\mathcal{T} = \{(u_1, \ldots, *_1, \ldots, u_2), (u_2, \ldots, *_2, \ldots, u_3), \ldots, (u_s, \ldots, *_s, \ldots, u_{s+1})\}$ with $x$ named nodes and star in the middle between $u_i$ and $u_{i+1}$ (assume $x$ to be even, $x$

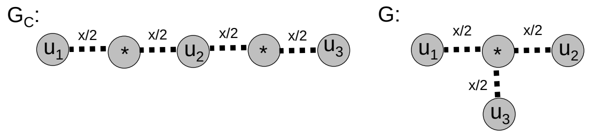


**Fig. 6** Estimation error for diameter.

does not include $u_i$ and $u_{i+1}$). It holds that $\textsc{Diam}(G_C) = s \cdot (x+2)$ whereas in a graph $G$ where all stars are merged, $\textsc{Diam}(G) = x+2$. There are $n = s(x+1)$ non-anonymous nodes, so $x = n/s - 1$. Figure 6 depicts an example. □

**Theorem 8** *For the maximal node degree* $\textsc{Deg}$*, we have* $\textsc{Deg}(G_1) - \textsc{Deg}(G_2) \leq 2(s - \gamma(G_*))$ *and* $\textsc{Deg}(G_1)/\textsc{Deg}(G_2) \leq s - \gamma(G_*) + 1$*. There are instances* $G_1, G_2$ *that reach these bounds.*

*Proof* Each occurrence of a node in a trace describes at most two links incident to this node. For the degree difference we only have to consider the links incident to at least one anonymous node, as the number of links between non-anonymous nodes is the same in $G_1$ and $G_2$. If all anonymous nodes can be merged into $\gamma(G_*)$ nodes in $G_1$ and all anonymous nodes are separate in $G_2$ the difference in the maximum degree is thus at most $2(s - \gamma(G_*))$, as there can be at most $s - \gamma(G_*) + 1$ nodes merged into one node and the minimal maximum degree of a node in $G_2$ is two. This bound is tight, as the trace set $T_i = \{v_i, *, w_i\}$ for $1 \leq i \leq s$ containing $s$ stars can be represented by a graph with one anonymous node of degree $2s$ or by a graph with $s$ anonymous nodes of degree two each. For the ratio of the maximal degree we can ignore links between non-anonymous nodes as well, as these only decrease the ratio. The highest number of links incident at node $v$ with one endpoint in the set of anonymous nodes is $s - \gamma(G_*) + 1$ for non-anonymous nodes and $2(s - \gamma(G_*) + 1)$ for anonymous nodes, whereas the lowest number is two. □

The number of triangles is another important topology characteristic that indicates how well meshed a network is.

**Theorem 9** *Let* $C_3(G)$ *be the number of cycles of length* 3 *of the graph G. It holds that* $C_3(G_1) - C_3(G_2) \leq 2s(s-1)$*, which can be reached. The relative error* $C_3(G_1)/C_3(G_2)$ *can be arbitrarily large unless the number of links between non-anonymous nodes exceeds* $n^2/4$ *in which case the ratio is upper bounded by* $2s(s-1)+1$.

*Proof Upper bound:* Each node which is part of a triangle has at least two incident edges. Thus, a node $v$ can be part of at most $\binom{\textsc{Deg}(v)}{2}$ triangles, where $\textsc{Deg}(v)$ denotes $v$'s degree. As a consequence the number of triangles containing an anonymous node in an inferrable topology with $a$ anonymous nodes $u_1, \ldots u_a$ is at most $\sum_{j=1}^{a} \binom{\textsc{Deg}(u_j)}{2}$. Given $s$, this sum is maximized if $a = 1$ and $\textsc{Deg}(u_1) = 2s$ as $2s$ is the maximum degree possible due to Theorem 8. Thus there can be at most $s \cdot (2s-1)$ triangles containing an anonymous node in $G_1$. The number of triangles with at least one anonymous node is minimized in $G_C$ because in the canonic graph the degrees of the anonymous nodes are minimized, i.e, they are always exactly two. As a consequence there cannot be more than $s$ such triangles in $G_C$.

If the number of such triangles in $G_C$ is smaller by $x$, then the number of of triangles with at least one anonymous node in the topology $G_1$ is upper bounded by $s \cdot (2s-1) - x$. The difference between the triangles in $G_1$ and $G_2$ is thus at most $s(2s-1) - x - s + x = 2s(s-1)$.

*Example meeting this bound:* If the non-anonymous nodes form a complete graph and all star nodes can be merged into one node in $G_1$ and $G_2 = G_C$, then the difference in the number of triangles matches the upper bound. Consequently it holds for the ratio of triangles with anonymous nodes that it does not exceed $(s(2s-1)-x)/(s-x)$. Thus the ratio can be infinite, as $x$ can reach $s$. However, if the number of links between $n$ non-anonymous nodes exceeds $n^2/4$ then there is at least one triangle, as the densest complete bipartite graph contains at most $n^2/4$ links. □

## 4 Full Exploration

So far, we assumed that the trace set $\mathcal{T}$ contains each node and link of $G_0$ at least once. At first sight, this seems to be the best we can hope for. However, sometimes traces exploring the vicinity of anonymous nodes in more than one trace yields additional information that help to characterize $\mathcal{G}_\mathcal{T}$ better.

This section introduces the concept of *fully explored networks*: A trace set $\mathcal{T}$ fully explores a network if contains sufficiently many traces such that the distances between non-anonymous nodes can be estimated accurately.

**Definition 7 (Fully Explored Topologies)** A topology $G_0$ is fully explored by a trace set $\mathcal{T}$ if it contains all nodes and links of $G_0$ and for each pair $\{u, v\}$ of non-anonymous nodes in the same component of $G_0$ there exists a trace $T \in \mathcal{T}$ containing both nodes $u \in T$ and $v \in T$.

A trace set for a fully explored network is the optimal input for generic topology inference in the sense that aspects that cannot be inferred well in a fully explored topology model are infeasible to infer without additional assumptions on $G_0$. Put differently, a fully exploring set of traces has the property that adding any additional traces does not change the set of topologies which are consistent with the given traces. Thus, this section provides "upper bounds" on what can be learned from topology inference. However, in the following, we will make the simplifying assumption that routing occurs along shortest paths only ($\alpha = 1$).

Let us again study the properties of the family of inferrable topologies fully explored by a trace set. Obviously, all the upper bounds from Section 3 are still valid for fully explored topologies. In the following, let $G_1, G_2 \in \mathcal{G}_{\mathcal{T}}$ be arbitrary representatives of $\mathcal{G}_{\mathcal{T}}$ for a fully explored trace set $\mathcal{T}$. A direct consequence of the Definition 7 concerns the number of connected components and the stretch. (Recall that the stretch is defined with respect to named nodes only, and since $\alpha = 1$, a 1-consistent inferrable topology cannot include a shorter path between $u$ and $v$ than the one that must appear in a trace of $\mathcal{T}$.)

**Corollary 1** *It holds that* $\text{COMP}(G_1) = \text{COMP}(G_2)$ *(=* $\text{COMP}(G_0)$*) and the stretch is 1.*

The proofs for the following theorems are analogous to our former proofs, as the main difference is the fact that there might be more conflicts, i.e., edges in $G_*$.

**Theorem 10** *For fully explored networks it holds that* $|V_1| - |V_2| \leq s - \gamma(G_*) \leq s - 1$ *and* $|V_1|/|V_2| \leq (n+s)/(n+\gamma(G_*)) \leq (2+s)/3$*. Moreover,* $|E_1| - |E_2| \in 2(s-\gamma(G_*))$ *and* $|E_1|/|E_2| \leq (\nu+2s)/(\nu+2) \leq s$*, where* $\nu$ *denotes the number of links between non-anonymous nodes. There are traces with inferrable topologies* $G_1, G_2$ *reaching these bounds.*

**Theorem 11** *For the maximal node degree, we have* $\text{DEG}(G_1) - \text{DEG}(G_2) \leq 2(s - \gamma(G_*))$ *and* $\text{DEG}(G_1)/\text{DEG}(G_2) \leq s - \gamma(G_*) + 1$*. There are instances* $G_1, G_2$ *that reach these bounds.*

*Proof* The proof for the upper bound is analogous to the the proof for arbitrary trace sets. To prove that this bound can be reached for fully explored networks, we need to add traces to the trace set to ensure that all pairs of named nodes appear in the trace but does not change the degrees of anonymous nodes. To this end we add a named node $u$ for each pair $\{v, w\}$ that is not in the trace set yet to $G_0$ and a trace $T = \{v, u, w\}$. This does not increase the maximum degree and guarantees full exploration. □

From Corollary 1 we know that fully explored scenarios yield a perfect stretch of one. However, regarding the diameter, the situation is different since distances between anonymous nodes play a role.

**Theorem 12** *Let* $\text{DIAM}$ *denote the graph diameter and* $\text{DIAM}(G_1) > \text{DIAM}(G_2)$*. For connected topologies* $G_1, G_2$ *it holds that* $\text{DIAM}(G_1)/\text{DIAM}(G_2) \leq 2$*, and there are instances* $G_1, G_2$ *that reach this bound. Moreover, it holds that* $\text{DIAM}(G_1) - \text{DIAM}(G_2) \leq s/2$*, and there are instances with* $\text{DIAM}(G_1) - \text{DIAM}(G_2) = s/2$*.*

*Proof* We first prove the upper bound for the relative case. Note that the maximal distance between two anonymous nodes $\text{MAP}(*_1)$ and $\text{MAP}(*_2)$ in an inferred topology component cannot be larger than twice the distance of two named nodes $u$ and $v$: from Definition 7 we know that there must be a trace in $\mathcal{T}$ connecting $u$ and $v$, and the maximal distance $\delta$ of a pair of named nodes is given by the path of the trace that includes $u$ and $v$. Therefore, and since any trace starts and ends with a named node, any star can be at a distance at a distance $\delta/2$ from a named node. Therefore, the maximal distance between $\text{MAP}(*_1)$ and $\text{MAP}(*_2)$ is $\delta/2+\delta/2$ to get to the corresponding closest named nodes, plus $\delta$ for the connection between the named nodes. As according to Corollary 1, the distance between named nodes is the same in all inferred topologies, the diameter of inferred topologies can vary at most by a factor of two.

We now construct an example that reaches this bound. Consider a topology consisting of a center node $c$ and four rays of length $k$. Let $u_1, u_2, u_3, u_4$ be the "end nodes" of each ray. We assume that all these nodes are named. Now add two chains of anonymous nodes of length $2k+1$ between nodes $u_1$ and $u_2$, and between nodes $u_3$ and $u_4$ to the topology. The trace set consists of the minimal trace set to obtain a fully explored topology: six traces of length $2k+1$ between each pair of end nodes $u_1, u_2, u_3, u_4$. Now we add two traces of length $2k+1$ between nodes $u_1$ and $u_2$, and between nodes $u_3$ and $u_4$. These traces explore the anonymous chains and have the following shape: $T_7 = (u_1, *_1, \ldots, *_k, \sigma, *_{k+1}, \ldots, *_{2k}, u_2)$ and $T_8 = (u_3, *_{2k+1}, \ldots, *_{3k}, \sigma', *_{3k+1}, \ldots, *_{4k}, u_4)$, where $\sigma$ and $\sigma'$ are stars. Let $G_1 = G_C$ and $G_2$ be the inferrable graph where $\sigma$ and $\sigma'$ are merged. The resulting diameters are $\text{DIAM}(G_1) = 4k+2$ and $\text{DIAM}(G_2) = 2k+1$. Since $s = 4k+2$, the difference can thus be as large as $s/2$. Note that this construction also yields the bound of the relative difference: $\text{DIAM}(G_1)/\text{DIAM}(G_2) = (4k+2)/(2k+1) = 2$.

Finally, it remains to prove the absolute upper bound. For the sake of contradiction, let us assume that $\text{DIAM}(G_1) - \text{DIAM}(G_2) > s/2$. Let $u, v$ be two nodes whose distance in $G_1$ constitutes $G_1$'s diameter, i.e., $\text{DIAM}(G_1) = d_{G_1}(u, v)$. Since we attend to a fully exploring trace set $\mathcal{T}$, $\text{DIAM}(G_1) \neq \text{DIAM}(G_2)$ implies that at least one of the nodes $u, v$ must be anonymous. We first assume that $u$ is a named node, and $v$ is anonymous; the case where $u$ is anonymous and $v$ is non-anonymous is symmetric. We define $*_v = v$. Let $T_u \in \mathcal{T}$ be the trace that includes $*_v$, and let $w$ be the closest named node to $*_v$ in $T_u$, i.e., $w = \arg\min_{x \in \mathcal{ID}} d_{T_u}(x, *_v)$. Since fully exploring trace sets yield unique distances between named nodes if $\alpha = 1$, $d_{G_1}(u, w) = d_{G_2}(u, w)$ and $d_{G_1}(u, w) + d_{G_1}(w, *_v) \geq d_{G_1}(u, *_v)$ due to the triangle inequality. With the definition of the diameter ($\text{DIAM}(G_2) = \max_{x,y} d_{G_2}(x, y)$), we obtain $\text{DIAM}(G_2) + d_{G_1}(w, *_v) \geq d_{G_1}(u, w) + d_{G_1}(w, *_v) \geq d_{G_1}(u, *_v) = \text{DIAM}(G_1) > s/2 + \text{DIAM}(G_2)$ and therefore

$d_{G_1}(w, *_v) > s/2$. However, since $s$ is the total number of stars in $\mathcal{T}$, we have that $d_{G_1}(w, *_v) \leq \lfloor s/2 \rfloor$ which yields the desired contradiction.

It remains to consider the case where both $u$ and $v$ are anonymous. Let us call the corresponding nodes in $G_1$ be $*_u$ and $*_v$. We assume that $d_{G_1}(*_u, *_v) > s/2 + \textsc{Diam}(G_2)$. Let $T_u$ and $T_v$ be the traces containing $*_u$ and $*_v$, and let $x_u = \arg\min_{x \in \mathcal{ID}} d_{T_u}(x, *_u)$ and $x_v = \arg\min_{x \in \mathcal{ID}} d_{T_v}(x, *_v)$ the closest non-anonymous nodes. (If $*_u$ and $*_v$ occur in the same trace, the longest distance is assumed between named nodes.) We have that $d_{G_1}(*_u, x_u) + d_{G_1}(*_v, x_v) + d_{G_1}(x_u, x_v) \geq d_{G_1}(*_u, *_v)$ due to the triangle inequality. Therefore, $d_{G_1}(*_u, x_u) + d_{G_1}(*_v, x_v) > s/2$. Let $s_u$ denote the number of stars in $T_u$ and let $s_v$ be the number of stars in $T_v$. It holds that $d_{G_1}(*_u, x_u) \leq \lfloor (s_u - 1)/2 \rfloor$ and $d_{G_1}(*_v, x_v) \leq \lfloor (s_v - 1)/2 \rfloor$, so $s/2 < d_{G_1}(*_u, x_u) + d_{G_1}(*_v, x_v) \leq \lfloor (s_u - 1)/2 \rfloor + \lfloor (s_v - 1)/2 \rfloor \leq s/2$. Contradiction. □

The number of triangles with anonymous nodes can still not be estimated accurately in the fully explored scenario.

**Theorem 13** *There exist graphs where* $C_3(G_1) - C_3(G_2) = s(s-1)/2$*, and the relative error* $C_3(G_1)/C_3(G_2)$ *can be arbitrarily large.*

*Proof* Given the number of stars $s$, we construct a trace set $\mathcal{T}$ with two inferrable graphs such that in one graph the number of triangles with anonymous nodes is $s(s-1)/2$ and in the other graph there are no such triangles. As a first step we add $s$ traces $T_i = (v_i, *_i, w)$ to the trace set $\mathcal{T}$, where $1 \leq i \leq s$. To make this trace set fully explored we add traces for each pair $v_i, v_j$ to $\mathcal{T}$ as a second step, i.e., traces $T_{i,j} = (v_i, v_j)$ for $1 \leq i \leq s$ and $1 \leq j \leq s$. The resulting trace set contains $s$ stars and none of the stars are in conflict with each other. Thus the graph $G_1$ merging all stars into one anonymous node is inferrable from this trace and the number of triangles where the anonymous node is part of is $s(s-1)/2$. Let $G_2$ be the canonic graph of this trace set. This graph does not contain any triangles with anonymous nodes and hence the difference $C(G_1) - C(G_2)$ is $s(s-1)/2$.

To see that the ratio can be unbounded look at the trace set $\{(v, *_1, w), (u, *_2, w), (u, v)\}$. This set is fully explored since all pairs of named nodes appear in a trace. The graph where the two stars are merged has one triangle and the canonic graph has no triangle. □

## 5 Conclusion

Our work is to be viewed as a first step to shed light onto the similarity of inferrable topologies if the trace sets are based on very basic axioms and no assumptions on the underlying network are taken (e.g., assumptions on power-law properties of the degree distribution). In other words, we consider the worst case: arbitrary networks. Using our formal framework we show that the topologies for a given trace set may differ significantly. Thus, it is impossible to accurately characterize topological properties of complex networks. Note that while this is a negative result from the perspective of application designers who may want to exploit the topology, certain players in the Internet (e.g., ISPs) are interested in keeping their topology as a business secret. [25]

To complement the general analysis, we propose the notion of fully explored networks or trace sets, as a "best possible scenario". As expected, we find that fully exploring traces allow us to determine several properties of the network more accurately; however, it also turns out that even in this scenario, other topological properties are inherently hard to compute. Our results are summarized in Figure 7.

Our work opens several directions for future research. So far we have only investigated fully explored networks with short path routing ($\alpha = 1$), and a scenario with suboptimal routes might lead to different results. One may also study whether the minimal inferrable topologies considered in, e.g., [1,2], share more similarities than the whole set of inferrable graphs. More importantly, while this article merely presented bounds for the general worst-case, it is of great interest to devise (efficient) algorithms that compute, for a *given trace set*, worst-case bounds for the properties under consideration. For example, such approximate bounds would be helpful to decide whether additional measurements are needed. Moreover, such algorithms may even give advice on the locations at which such measurements would be most useful. Finally, it would also be interesting to study whether the additional knowledge that a network must belong to a certain graph family (e.g., types of backbone networks) may render inference more efficient and accurate.

### Acknowledgments

Stefan Schmid would like to thank H. B. Acharya and Ingmar Poese for initial discussions in Calcutta and Berlin. We are also grateful to Bernhard Ager, Georgios Smaragdakis and Steve Uhlig for their comments and the literature pointers.

### References

1. H. Acharya and M. Gouda. The weak network tracing problem. In *Proc. Int. Conference on Distributed Computing and Networking (ICDCN)*, pages 184–194, 2010.
2. H. Acharya and M. Gouda. On the hardness of topology inference. In *Proc. Int. Conference on Distributed Computing and Networking (ICDCN)*, pages 251–262, 2011.
3. M. Acharya, H. a nd Gouda. A theory of network tracing. In *Proc. 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, pages 62–74, 2009.

| Property/Scenario | Arbitrary | | Fully Explored ($\alpha = 1$) | |
|---|---|---|---|---|
| | $G_1 - G_2$ | $G_1/G_2$ | $G_1 - G_2$ | $G_1/G_2$ |
| # of nodes | $\leq s - \gamma(G_*)$ | $\leq (n+s)/(n+\gamma(G_*))$ | $\leq s - \gamma(G_*)$ | $\leq (n+s)/(n+\gamma(G_*))$ |
| # of links | $\leq 2(s - \gamma(G_*))$ | $\leq (\nu + 2s)/(\nu + 2)$ | $\leq 2(s - \gamma(G_*))$ | $\leq (\nu + 2s)/(\nu + 2)$ |
| # of connected components | $\leq n/2$ | $\leq n/2$ | $= 0$ | $= 1$ |
| Stretch | - | $\leq (N-1)/2$ | - | $= 1$ |
| Diameter | $\leq (s-1)/s \cdot (N-1)$ | $\leq s$ | $\leq s/2$ | $\leq 2$ |
| Max. Deg. | $\leq 2(s - \gamma(G_*))$ | $\leq s - \gamma(G_*) + 1$ | $\leq 2(s - \gamma(G_*))$ | $\leq s - \gamma(G_*) + 1$ |
| Triangles | $\leq 2s(s-1)$ | $\infty$ | $\leq 2s(s-1)/2$ | $\infty$ |

**Fig. 7** Summary of our bounds on the properties of inferrable topologies. $s$ denotes the number of stars in the traces, $n$ is the number of named nodes, $N = n + s$, and $\nu$ denotes the number of links between named nodes. Note that trace sets meeting these bounds exist for all properties for which we have upper bounds.

4. D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 694–703, 2005.
5. B. Ager, N. Chatzis, A. Feldmann, N. Sarrar, S. Uhlig, and W. Willinger. Anatomy of a large european ixp. In *Proc. ACM SIGCOMM*, 2012.
6. N. Alon, Y. Emek, M. Feldman, and M. Tennenholtz. Economical graph discovery. In *Proc. 2nd Symposium on Innovations in Computer Science (ICS)*, 2011.
7. A. Anandkumar, A. Hassidim, and J. Kelner. Topology discovery of sparse random graphs with few participants. In *Proc. SIGMETRICS*, 2011.
8. B. Augustin, X. Cuvellier, B. Orgogozo, F. Viger, T. Friedman, M. Latapy, C. Magnien, and R. Teixeira. Avoiding traceroute anomalies with paris traceroute. In *Proc. 6th ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 153–158, 2006.
9. P. Barford, A. Bestavros, J. Byers, and M. Crovella. On the marginal utility of network topology measurements. In *Proc. 1st ACM SIGCOMM Workshop on Internet Measurement (IMW)*, pages 5–17, 2001.
10. A. Bestavros, J. W. Byers, and K. A. Harfoush. Inference and labeling of metric-induced network topologies. *IEEE Trans. Parallel Distrib. Syst.*, 16:1053–1065, 2005.
11. M. Buchanan. Data-bots chart the internet. *Science*, 813(3), 2005.
12. B. Cheswick, H. Burch, and S. Branigan. Mapping and visualizing the internet. In *Proc. USENIX Annual Technical Conference (ATEC)*, 2000.
13. L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. Exploring networks with traceroute-like probes: Theory and simulations. *Theoretical Computer Science*, 355(1):6–24, 2006.
14. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. SIGCOMM*, pages 251–262, 1999.
15. P. Gill, M. Arlitt, Z. Li, and A. Mahanti. The flattening internet topology: Natural evolution, unsightly barnacles or contrived collapse? *Passive and Active Network Measurement*, pages 1–10, 2008.
16. M. Gunes and K. Sarac. Resolving anonymous routers in internet topology measurement studies. In *Proc. INFOCOM*, 2008.
17. K. Harfoush, A. Bestavros, and J. Byers. Measuring capacity bandwidth of targeted path segments. *IEEE/ACM Trans. Netw.*, 17:80–92, 2009.
18. X. Jin, W.-P. Yiu, S.-H. Chan, and Y. Wang. Network topology inference based on end-to-end measurements. *IEEE Journal on Selected Areas in Communications*, 24(12):2182 –2195, 2006.
19. R. Khalili, N. Gast, M. Popovic, U. Upadhyay, and J.-Y. Le Boudec. MPTCP is not pareto-optimal: performance issues and a possible solution. In *Proc. CoNEXT*, pages 1–12, 2012.
20. C. Labovitz, A. Ahuja, S. Venkatachary, and R. Wattenhofer. The impact of internet policy and topology on delayed routing convergence. In *Proc. 20th IEEE INFOCOM*, volume 1, pages 537–546, 2001.
21. A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. In *Proc. 22nd IEEE INFOCOM*, volume 1, pages 332 – 341, 2003.
22. S. Paul, K. K. Sabnani, J. C. Lin, and S. Bhattacharyya. Reliable multicast transport protocol (rmtp). *IEEE Journal on Selected Areas in Communications*, 5(3), 1997.
23. V. Paxson. End-to-end routing behavior in the internet. In *ACM SIGCOMM Computer Communication Review*, volume 26, pages 25–38. ACM, 1996.
24. Y.-A. Pignolet, G. Tredan, and S. Schmid. Adversarial VNet embeddings: A threat for ISPs? In *INFOCOM Mini-Conference*, 2013.
25. I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving content delivery using provider-aided distance information. In *Proc. ACM IMC*, 2010.
26. H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin. The impact of routing policy on internet paths. In *Proc. 21st IEEE INFOCOM*, volume 2, pages 736–742, 2002.
27. X. Xiao, A. Hannan, B. Bailey, and L. Ni. Traffic engineering with mpls in the internet. *Network, IEEE*, 14(2):28 –33, 2000.
28. B. Yao, R. Viswanathan, F. Chang, and D. Waddington. Topology inference in the presence of anonymous routers. In *Proc. 22nd IEEE INFOCOM*, pages 353–363, 2003.