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

Yvonne Anne Pignolet[1], Stefan Schmid[2], and Gilles Tredan[2]

[1] ABB Research, Switzerland
yvonne-anne.pignolet@ch.abb.com
[2] TU Berlin & T-Labs, Germany
{stefan,gilles}@net.t-labs.tu-berlin.de

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 paper 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.

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.

Since knowledge of the network characteristics is crucial for many applications (e.g., *RMTP* [13], or *PaDIS* [14]), the research community implements measurement

* Due to space constraints, some proofs are omitted in this document. They are available from the ArXiv document server (ID: 1105.5236).

tools to analyze at least the main properties of the network. The results can then, e.g., be used to design more efficient network protocols in the future.

This paper 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 paper 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 [9] 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. This 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 [9] was published, many additional results have been obtained, e.g., by conducting a distributed computing approach to increase the number of measurement points [7]. However, our understanding 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 paper.

In the field of *network tomography*, 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 [5]. For a good discussion of this approach as well as results for a routing model along shortest and second shortest paths see [5]. For example, [5] shows that for sparse random graphs, a relatively small number of cooperating participants is sufficient to discover a network fairly well.

The classic tool to discover Internet topologies is traceroute [8]. Unfortunately, there are several problems with this approach that render topology inference difficult, such as *aliasing* or *load-balancing*, which has motivated researchers to develop new tools such as *Paris Traceroute* [6, 11]. Another complication stems from the fact that routers may appear as stars in the trace since they are overloaded or since they are configured not to send out any ICMP responses. The lack of complete information in the trace set renders the accurate characterization of Internet topologies difficult.

This paper attends to the problem of anonymous nodes and assumes a conservative, "worst-case" perspective that does not rely on any assumptions on the underlying network. There are already several works on the subject. Yao et al. [16] initiated the study of possible candidate topologies for a given trace set and suggested computing the *minimal topology*, that is, the topology with the minimal number of anonymous nodes, which turns out to be NP-hard. Consequently, different heuristics have been proposed [10, 11].

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 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]. As proved in [2], the problem is still hard (in the sense that there are many minimal networks corresponding to a trace set), even with only two anonymous nodes, symmetric routing and without aliasing.

In contrast to this line of research on cardinalities, we are interested in the *network properties*. If the inferred topologies share the most important characteristics, the negative results 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.

Finally, in a broader context, Alon et al. [4] recently initiated the study of the multi-agent exploration of *link weights* in known network topologies, and derived bounds on the number of rounds and the number of agents required to complete the discovery of the edge weights or a shortest path.

1.2 Our Contribution

This paper 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 initiate 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 paper 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 there are situations where there indeed exist so many inferrable topologies. Especially, there are problem instances where the cardinality of the set of inferrable topologies equals the *Bell number*. This insight complements (and generalizes to arbitrary, not only minimal, inferrable topologies) existing cardinality results.

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

2 Model

Let $\mathcal{T}$ denote the set of traces obtained from probing (e.g., by traceroute) a (not necessarily connected and undirected) network $G_0 = (V_0, E_0)$ with *nodes* or *vertices* V_0 (the set of routers) and *links* or *edges* E_0. We assume that G_0 is static during the probing time (or that probing is instantaneous). 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. Therefore, each trace $T \in \mathcal{T}$ 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): Σ 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 $\mathcal{T}$.

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 σ_1 and σ_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 [15] and on the convergence time [12].

AXIOM 2 (*α-(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 σ_1, σ_2 in trace T, then the shortest path connecting the two (named or anonymous) nodes corresponding to σ_1 and σ_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.)

A natural axiom to merge traces is the following.

AXIOM 3 (*Trace Merging*): For two traces $T_1, T_2 \in \mathcal{T}$ for which $\exists \sigma_1, \sigma_2, \sigma_3$, where σ_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 σ_1 and σ_2, respectively, in G_0, is at most $d_{G_0}(\sigma_1, \sigma_3) \leq i + j$.

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 (α-consistently)* inferrable *from a trace set $\mathcal{T}$ if axioms* AXIOM *0,* AXIOM *1,* AXIOM *2 (with parameter α), and* AXIOM *3 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, AXIOM 2 and AXIOM 3 are fulfilled by definition. However, there are instances where an α-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 α-consistent trace set $\mathcal{T}$. In the remainder of this paper, we will only consider settings where $G_0 \in \mathcal{G}_\mathcal{T}$.

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 α specifying the assumed routing consistency. Essentially, the goal of any topology inference algorithm ALG is to compute a mapping of the symbols Σ (appearing in $\mathcal{T}$) to nodes in an inferred topology G; or, in case the input parameters α 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.

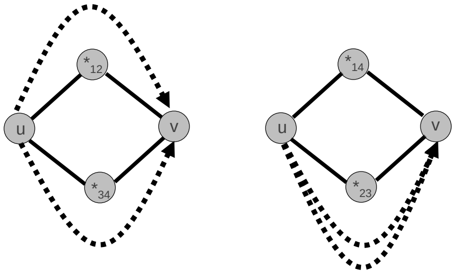

Fig. 1. Two non-isomorphic inferred topologies, i.e., different mapping functions lead to these topologies.

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 paper, 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 1: 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 paper, 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 Σ, 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(σ) *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 β 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(σ) to refer to a symbol σ's corresponding node in G_0.

In the remainder of this paper, we will often assume that AXIOM 0 is given. Moreover, note that AXIOM 3 is redundant. Therefore, in our proofs, we will not explicitly cover AXIOM 0, and it is sufficient to show that AXIOM 1 holds to prove that AXIOM 3 is satisfied.

Lemma 1. AXIOM *1 implies* AXIOM *3.*

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 π_1 of length at most i between the nodes corresponding to σ_1 and σ_2 in G and a path π_2 of length at most j between the nodes corresponding to σ_2 and σ_3 in G. The combined path can only be shorter, and hence the claim follows. □

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., σ_1 follows after σ_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 σ_1 and σ_2.*

Note that G_C is indeed an inferrable topology. In this case, $\text{MAP} : \Sigma \rightarrow \Sigma$ is the identity function.

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

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 $\mathcal{T}$.

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., $\text{MAP}(*_1) \neq \text{MAP}(*_2)$, 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 $\text{MAP}(*_1)$ 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 $\text{MAP}(*_1) = \text{MAP}(*_2)$ 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 $\text{MAP}(*_1) = \text{MAP}(*_2)$ 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 α-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 α.

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

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 γ 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 γ 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_γ is defined as the topology* induced *by γ. G_γ describes the graph G_C where nodes of the same color are contracted: two vertices u and v represent the same node in G_γ, i.e.,* $\textsc{Map}(*_i) = \textsc{Map}(*_j)$, *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 γ be a proper coloring of G_*. The coloring induced topology G_γ is a topology fulfilling* AXIOM *2 with a routing consistency of α', for some positive α'.*

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_*.*

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

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

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

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. □

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. *For any $\alpha > 0$, there is a trace set for which the number of non-isomorphic colorings of G_* equals $|\mathcal{G}_\mathcal{T}|$, in particular $|\mathcal{G}_\mathcal{T}| = B_s$, where $\mathcal{G}_\mathcal{T}$ is the set of inferrable and α-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). □

3.2 Properties

Even if the number of inferrable topologies is large, topology inference can still be useful if one is mainly interested in the 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 $\mathcal{G}_\mathcal{T}$ are close to G_0, there may be an option to conduct an efficient sampling analysis on random representatives. Therefore, in the following, we will take a closer look how much the members of $\mathcal{G}_\mathcal{T}$ differ.

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, and so on. 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 $\mathcal{G}_\mathcal{T}$.

As one might expect, the graph size can be estimated quite well.

Lemma 6. *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 ν denotes the number of edges between non-anonymous nodes. There are traces with inferrable topology G_1, G_2 reaching these bounds.*

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 $\mathcal{G}_\mathcal{T}$.

Lemma 7. *Let $\textsc{Comp}(G)$ denote the number of connected components of a topology G. Then, $|\textsc{Comp}(G_1) - \textsc{Comp}(G_2)| \leq n/2$. There are traces with inferrable topology G_1, G_2 reaching these bounds.*

An important criterion for topology inference regards 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 *ρ: $\rho = \max_{u,v \in \mathcal{ID}(G_0)} \max\{d_{G_0}(u,v)/d_G(u,v), d_G(u,v)/d_{G_0}(u,v)\}$.*

From Lemma 7 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 can be much longer or shorter than in G_0.

Figure 2 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 2 *left*), whereas the other topology merges the two anonymous nodes (Figure 2 *right*). The distances between v and w are $2(k + 2)$ and 2, respectively, implying a stretch of $k + 2$.

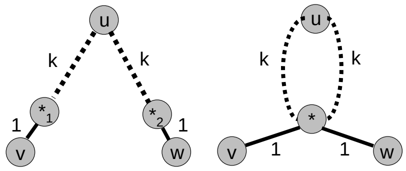

Fig. 2. Due to the lack of a trace between v and w, the stretch of an inferred topology can be large.

Lemma 8. *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 $\rho \leq (N - 1)/2$. There are traces with inferrable topology G_1, G_2 reaching these bounds.*

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

Lemma 9. *For connected topologies G_1, G_2 it holds that* $\textsc{Diam}(G_1) - \textsc{Diam}(G_2) \leq (s-1)/s \cdot \textsc{Diam}(G_C) \leq (s-1)(N-1)/s$ *and* $\textsc{Diam}(G_1)/\textsc{Diam}(G_2) \leq s$*, where* $\textsc{Diam}$ *denotes the graph diameter and* $\textsc{Diam}(G_1) > \textsc{Diam}(G_2)$*. 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 π be a longest path between two nodes u and v in G_C. In the extreme case, π is the only path determining the network diameter and π contains all star nodes. Then, the graph where all s stars are merged into one anonymous node has a minimal diameter of at least $\textsc{Diam}(G_C)/s$.

Example which meets 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 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 $x = (n-s-1)/s$. Figure 3 depicts an example. □

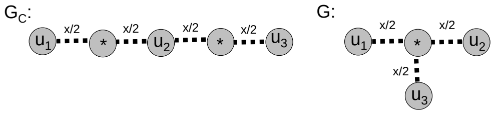

Fig. 3. Estimation error for diameter.

Lemma 10. *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.*

Another important topology measure that indicates how well meshed the network is, is the number of triangles.

Lemma 11. *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 $2s(s-1)+1$.*

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 different ways yields additional information that help to characterize $\mathcal{G}_\mathcal{T}$ better.

This section introduces the concept of *fully explored networks*: $\mathcal{T}$ 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 $v \in T$.*

In some sense, a trace set for a fully explored network is the best we can hope for. Properties that cannot be inferred well under the fully explored topology model are infeasible to infer without additional assumptions on G_0. In this sense, this section provides upper bounds on what can be learned from topology inference, and accordingly, we will constrain ourselves to routing 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 $\mathcal{T}$.)

Lemma 12. *It holds that $\text{COMP}(G_1) = \text{COMP}(G_2)$ ($= \text{COMP}(G_0)$) and the stretch is 1.*

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

Lemma 13. *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 ν denotes the number of links between non-anonymous nodes. There are traces with inferrable topology G_1, G_2 reaching these bounds.*

Lemma 14. *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.*

From Lemma 12 we know that fully explored scenarios yield a perfect stretch of one. However, regarding the diameter, the situation is different in the sense that distances between anonymous nodes play a role.

Lemma 15. *For connected topologies G_1, G_2 it holds that $\text{DIAM}(G_1)/\text{DIAM}(G_2) \leq 2$, where DIAM denotes the graph diameter and $\text{DIAM}(G_1) > \text{DIAM}(G_2)$. There are instances G_1, G_2 that reach this bound. Moreover, there are instances with $\text{DIAM}(G_1) - \text{DIAM}(G_2) = s/2$.*

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

Lemma 16. *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.*

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	$\leq (s-1)/s \cdot (N-1)$	$\leq s$	$s/2$ (¶)	2
Max. Deg.	$\leq 2(s - \gamma(G_*))$	$\leq s - \gamma(G_*) + 1$	$\leq 2(s - \gamma(G_*))$	$\leq s - \gamma(G_*) + 1$
Triangles	$\leq 2s(s-1)$	∞	$\leq 2s(s-1)/2$	∞

Fig. 4. 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 ν denotes the number of links between named nodes. Note that trace sets meeting these bounds exist for all properties for which we have tight or upper bounds. For the entry marked with (¶), only "lower bounds" are derived, i.e., examples that yield at least the corresponding accuracy; as the upper bounds from the arbitrary scenario do not match, how to close the gap remains an open question.

5 Conclusion

We understand our work as a first step to shed light onto the similarity of inferrable topologies based on most basic axioms and without any assumptions on power-law properties, i.e., in the worst case. 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. 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 4.

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 still needs to be investigated. One may also study whether the minimal inferrable topologies considered in, e.g., [1, 2], are more similar in nature. More importantly, while this paper presented results for the general worst-case, it would be interesting to devise 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, maybe such algorithms may even give advice on the locations at which such measurements would be most useful.

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 Steve Uhlig and Bernhard Ager for their comments and the literature pointers. A full version of this article including all proofs is available on ArXiv (ID: 1105.5236).

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. H. B. Acharya and M. G. Gouda. A theory of network tracing. In *Proc. 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, pages 62–74, 2009.
4. N. Alon, Y. Emek, M. Feldman, and M. Tennenholtz. Economical graph discovery. In *Proc. 2nd Symposium on Innovations in Computer Science (ICS)*, 2011.
5. A. Anandkumar, A. Hassidim, and J. Kelner. Topology discovery of sparse random graphs with few participants. In *Proc. SIGMETRICS*, 2011.
6. 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.
7. M. Buchanan. Data-bots chart the internet. *Science*, 813(3), 2005.
8. B. Cheswick, H. Burch, and S. Branigan. Mapping and visualizing the internet. In *Proc. USENIX Annual Technical Conference (ATEC)*, 2000.
9. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. SIGCOMM*, pages 251–262, 1999.
10. M. Gunes and K. Sarac. Resolving anonymous routers in internet topology measurement studies. In *Proc. INFOCOM*, 2008.
11. 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.
12. C. Labovitz, A. Ahuja, S. Venkatachary, and R. Wattenhofer. The impact of internet policy and topology on delayed routing convergence. In *Proc. 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2001.
13. 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.
14. I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving content delivery using provider-aided distance information. In *Proc. ACM IMC*, 2010.
15. H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin. The impact of routing policy on internet paths. In *Proc. INFOCOM*, volume 2, pages 736–742, 2002.
16. B. Yao, R. Viswanathan, F. Chang, and D. Waddington. Topology inference in the presence of anonymous routers. In *Proc. IEEE INFOCOM*, pages 353–363, 2003.