Bokeh: Obfuscating Physical Infrastructure Maps

Yugali Gullapalli
yugali@cs.wisc.edu
University of Wisconsin-Madison

Jeremy Koritzinsky
koritzinsky@cs.wisc.edu
University of Wisconsin-Madison

Meenakshi Syamkumar
ms@cs.wisc.edu
University of Wisconsin-Madison

Paul Barford
pb@cs.wisc.edu
University of Wisconsin-Madison

Ramakrishnan Durairajan
rkrish@cs.uoregon.edu
University of Oregon

Joel Sommers
jsommers@colgate.edu
Colgate University

ABSTRACT

Physical infrastructures that facilitate *e.g.,* delivery of power, water and communication capabilities are of intrinsic importance in our daily lives. Accurate maps of physical infrastructures are important for permitting, maintenance, repair and growth but can be considered a commercial and/or security risk. In this paper, we describe a method for obfuscating physical infrastructure maps that removes sensitive details while preserving key features that are important in commercial and research applications. We employ a three-tiered approach: tier 1 does simple location fuzzing, tier 2 maintains connectivity details but randomizes node/link locations, while at tier 3 only distributional properties of a network are preserved. We implement our tiered approach in a tool called *Bokeh* which operates on GIS shapefiles that include detailed location information of infrastructure and produces obfuscated maps. We describe a case study that applies Bokeh to a number of Internet Service Provider maps. The case study highlights how each tier removes increasing amounts of detail from maps. We discuss how Bokeh can be generally applied to other physical infrastructures or in local services that are increasingly used for e-marketing.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**.

LocalRec'19, November 5, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6963-3/19/11…$15.00
https://doi.org/10.1145/3356994.3365501

KEYWORDS

Geographic Information System; Shapefile Obfuscation; Network Map Obfuscation; Internet Physical Infrastructure Maps

1 INTRODUCTION

Accurate maps of the geographic characteristics and features of physical infrastructures such as power, water and communication systems are routinely generated and maintained by both public and private entities. These maps are important for a wide range of applications including inventory management, risk assessment, permitting, maintenance, repair and growth. These maps are also important in a wide variety of research contexts that seek to assess complex combinations of characteristics or capabilities and/or make fundamental and longer term improvements to infrastructure.

Our study is concerned with maps that convey characteristics of physical infrastructure. Such maps typically include details of locations of key components (*e.g.,* buildings), which we refer to as *nodes*, and connectivity between components (*e.g.,* various forms of rights of way, wires, conduits, etc.), which we refer to as *links*. Representations of such infrastructure appear as a graph or network on a map. Moreover, they are often instantiated in a standard GIS format such as ESRI's shapefile or Google's KML/KMZ so that they can be easily visualized, analyzed and combined with other data. We argue that there are two fundamental features of such maps that define their accuracy. *Location accuracy* is the correspondence between the represented geographic locations of individual nodes/links in a map with their actual (ground truth) locations. *Graph accuracy* is the correspondence between individual nodes and links in a map and their actual physical connectivity in the network. In each case an exact match between the map and the true physical infrastructure would mean the map is 100% accurate.

Unfortunately, there are risks in publishing 100% accurate physical maps. First is the risk of attack on the physical infrastructure locations identified in the maps. These attacks seek to damage nodes and/or links causing outages that could have broader impact. Examples of such attacks (which were

not necessarily enabled by the availability of infrastructure maps) include the as-yet-unsolved cable cuts in the greater San Francisco area in 2015 [12]. Second is the risk of "remote attack", which is best exemplified in communication infrastructure. Examples of these include denial of service, eavesdropping or other types of malicious activity that could be planned and facilitated remotely using detailed connectivity information [1, 36, 45]. Third is the risk of disclosing competitive information that may be used adversarially in the marketplace.

In this paper, we consider the problem of obfuscating physical infrastructure maps. The goals for our study are to *(i)* create a methodology and tool that enable maps of physical infrastructure to be obfuscated sufficiently to address the threats listed above and *(ii)* publish a set of obfuscated maps of infrastructure that can be used by the community for research. While much prior work has been done on the related issues of user anonymization in social networks and location-based applications, and historically, obfuscated (or deceptive) maps have been created for the purpose of concealing treasure [57], to the best of our knowledge this is the first study focused on obfuscation of infrastructure maps.

Our method for obfuscation assumes a base map in electronic form that includes both connectivity and location details (*i.e.,* latitude/longitude of nodes and links). Given our definition of accuracy, it is the representations of connectivity and location that are the target for obfuscation. Our approach, described in Section 2, is an amalgamation of techniques that are informed by prior work described in Section 5, and offers three *tiers* of obfuscation capabilities. Tier 1 does simple location fuzzing and produces maps with graph accuracy and minimal location obfuscation. Tier 2 also maintains graph accuracy but randomizes the locations of nodes/links thereby providing a higher level of location obfuscation. Tier 3 obfuscates both graph and location accuracy, and preserves only the distributional properties of the underlying graph. The algorithms at each tier provide configurable obfuscation of the base map. We have not completed a formal analysis of the resilience of our methodology to attackers who wish to deobfuscate maps, nor do we relate configurations to a specific accuracy metric. Those topics are discussed in more detail in Section 4 and are the subject of on-going efforts.

We implemented our method in a tool we call *Bokeh* [1], which is designed for simple and efficient application to standard GIS file formats. To demonstrate the efficacy of our methods, we conduct a case study in which we apply Bokeh to a set of Internet Service Provider maps from the Internet Atlas project [16] . Accurate maps of service provider networks (*i.e.,* co-location centers and fiber-optic links) and networked infrastructure (*e.g.*, data centers, cell towers, DNS servers, etc.) are important in a variety of research areas. Examples include the study of Internet configurations (*e.g.,* [15, 41]), opportunities for performance enhancement (*e.g.,* [7]), resilience to outages and attacks (*e.g.,* [19, 36, 52]), energy efficiency (*e.g.,* [10]), content delivery networks (*e.g.,* [56]), and future deployment opportunities (*e.g.,* [14, 30]). Such maps are also useful in commercial applications *e.g.,* for infrastructure maintenance and repair, and for advertising and sales. As we show in Section 3, each tier of obfuscation results in a map that is increasingly difficult to identify as the base map.

To summarize, this paper makes the following contributions. First, we identify and define the problem of physical infrastructure map obfuscation. Second, we develop and implement a three-tiered method for map obfuscation in a tool that we call Bokeh. Third, we demonstrate our method in a case study on a set of Internet Service Provider infrastructure maps from Internet Atlas [2].

2 OBFUSCATING LOCATION DETAILS

In this section we provide an overview of our physical infrastructure obfuscation framework. We discuss the framework's design and the obfuscation methods employed within it. Lastly, we describe how the framework has been implemented in Bokeh.

2.1 Overview

Our obfuscation framework is motivated by the need for researchers to use maps of physical infrastructure to evaluate capabilities, risks and new opportunities. Maps of physical infrastructure that embed sensitive geographic information are not typically released to the public or shared among researchers. Shapefiles [20] are a standard format for maps that can be viewed and analyzed and considered in combination with other data in Geographic Information Systems. They comprise a collection of geographic features like points and lines, that are generally used along with projection definitions to enable visualization of those features. While one approach to preparing a map for public release might be to expunge all geographic information, the resulting *utility* of the map would suffer, possibly rendering it useless for certain types of inquiry. At the same time, we recognize that map providers will have different levels of risk tolerance. For these reasons our framework is oriented around a set of *obfuscation tiers* that enable map providers to make explicit and configurable tradeoffs between the level of obfuscation and the potential research (or commercial) utility of the resulting map.

[1] In photography, *bokeh* refers to blurring out portions of an image in a way that enhances aesthetic quality—https://en.wikipedia.org/wiki/Bokeh.

[2] A set of obfuscated maps of 42 service provider networks is available from https://www.impactcybertrust.org. The Bokeh source code is available at https://github.com/yugaligullapalli/Bokeh.

2.2 Methodology

Our obfuscation framework is organized into 3 tiers, which heuristically relate to a range of obfuscation methods that might be employed to "fuzz" certain aspects—specifically, the location and graph accuracy features—of a network infrastructure map. As indicated in Table 1, at one end of the spectrum (tier 1) are methods designed to make small modifications to the precision and/or accuracy of geographic points. Maps obfuscated via methods at tier 1 retain the most utility for research and commercial activity, since connectivity information is preserved and location information is randomized *e.g.,* to city-level boundaries. On the other end of the spectrum (tier 3), aspects of location information and connectivity details may be completely removed or may be uniformly randomized. As a result, maps resulting from tier 3 methods are highly obfuscated but will have the lowest utility. We discuss algorithms and techniques employed at each tier in detail below.

2.2.1 Tier 1 Obfuscation. Tier 1, which offers minimal obfuscation, utilizes simple methods that *blur* or *fuzz* the precision and/or accuracy of geographic locations of nodes and links in a physical infrastructure map. The resulting obfuscated map can retain *e.g.,* city-level locations of nodes, accurate link distances and all node connectivity characteristics.

Advantages of approaches in this tier include the fact that precise locations (*e.g.*, street-level addresses) of nodes and links are no longer present, but the map still contains most of the original information, making it suitable for the broadest range of research or commercial activity. For example, in the context of Internet research, a map obfuscated at tier 1 would remain useful for risk analysis, performance analysis, robustness analysis, and network topology characteristics analysis, among other possibilities. Such maps would also be attractive for Internet Service Providers that want to advertise service options in relatively specific node locations (*e.g.,* areas in cities) and the full connectivity characteristics of their infrastructures (*e.g.,* https://www.zayo.com/solutions/global-network/). For commercial purposes, maps at tier 1 would remain also remain useful for distance and locality analyses for advertising, inventory management, etc.

A potential consequence of tier 1 approaches is that in rural areas with above-ground power or communications cabling or with very limited numbers of rights of way, it may still be possible to identify the exact locations of infrastructure. Thus, maps produced using methods at this tier provide the highest research utility with modest resilience to a malicious entity.

The specific methods at tier 1 include:

Geocentric Fuzzing: For geocentric fuzzing, we take a set of ellipsoid definitions for the earth, such as Delambre 1810, Andrae 1876, International 1909, as defined by the PROJ.4 library [51]. We randomly select one of these to be *ellipsoid Z*. For each latitude/longitude point in an input map, we randomly select one of the other ellipsoids, which we refer to as *ellipsoid A*, and project the point using ellipsoid A using the Universal Transverse Mercator (UTM) projection [27]. We then take the projected coordinate and project it back from UTM to a latitude/longitude coordinate using ellipsoid Z as the Earth ellipsoid. This method can fuzz each point anywhere from about 100 meters to 3000 meters away from its original position, depending on the set of projections used. Note that this method alters the *accuracy* of a point but retains its precision.

Accuracy Fuzzing: Given a fuzzing requirement p, which represents a *decimal digit position* in a latitude or longitude coordinate, replace all digits in the coordinate from position p to the end (to the right) with a value chosen uniformly at random. For example, if p is 0 (*i.e.*, the units place), randomly replacing digits from that position and lower will result in fuzzing of $\approx$ 111 km, whereas if the position is $p = -2$ (*i.e.*, the hundredths place), randomly replacing digits from that position and lower will result in fuzzing of $\approx$ 1.1 km.

Precision Truncation: For each point in an input map, we truncate the precision of the latitude and longitude components to d decimal places. For example, with 1 decimal place, a latitude/longitude coordinate provides precision of $\approx$ 11.1 km, whereas with 4 decimal places the precision is $\approx$ 11 meters. Note that this method reduces the *precision* of a point but retains its accuracy (at least up to the new level of precision). Truncating to a small number of decimal places may provide an appropriate level of obfuscation, making it unlikely that an adversary could identify the physical location of a given node or link. Moreover, for infrastructures that have equipment dispersed over a fairly localized area, this technique bears a resemblance to *k-anonymization* [53] since multiple nodes could potentially share the same coordinate in the obfuscated map. Thus this technique, while simple, has been proven to be effective in a variety of other contexts. For example, in the flow record data that Internet2 makes available to the research community, the low-order 11 bits of each IPv4 address are zeroed prior to data release [32].

2.2.2 Tier 2 Obfuscation. In tier 2, location features are pseudonymized using techniques akin to TCPdpriv [24], which provides *prefix-preserving* pseudonymization of IP addresses. The result is that points near to each other in an unmodified map of physical infrastructure remain near to each other in the obfuscated map, although the absolute location (*e.g.*, city) is changed. Because the structure is preserved, however, a determined adversary may be able to recover information about the original network. For example, if there is a set of 3 nodes close together but far away from any other nodes, this group of nodes may be reidentified in the

Table 1: Overview of network map obfuscation tiers in Bokeh.

	Description	Impact on map features
Tier 1 (least obfuscation)	Location features are *fuzzed*; detailed location information is no longer present.	Precision of geographic points (latitude/longitude) may be reduced; small random perturbations to the accuracy of geographic points may be performed; connectivity characteristics are preserved.
Tier 2	Location features are obfuscated via cryptographic methods.	Relative proximity of geographic points is typically preserved (*e.g.*, points within a city) as well as node connectivity characteristics.
Tier 3 (most obfuscation)	Location features are randomized or possibly removed.	Geographic information is removed or randomized without regard to preserving any location-based structure in the original map.

obfuscated map since the relative distances of those nodes will be preserved. Node connectivity characteristics are also unperturbed. Thus, it may be possible for an adversary to uncover original node or link locations by matching localized node and links structures in an obfuscated map with an unmodified map.

The specific methods at tier 2 include:

TCPdPriv-like obfuscation: We consider each coordinate for each point within a map as a pair of $X.Y$ latitude/longitude values, where X is the integral portion and Y is the decimal portion of the latitude or longitude. For the integral component X we compute a pseudorandom permuted value W with the constraint that latitude is constrained to the range $[-90, 89]$ and longitude is constrained to the range $[-180, 179]$. The value X might be thought of as a *prefix*, which is preserved through a pseudorandom permutation function. We then generate a random value Z and combine with W to give the coordinate value $W + 0.Z$.

Cryptographic obfuscation: We adapt the technique of Xu *et al.* [62] to obfuscate latitude/longitude values. We first shift values so that they are positive then scale them so that we operate on non-negative integers. (Our scaling preserves the first three decimal places, which limits precision to approximately 110 meters.) After applying the method of [62] to a given shifted and scaled latitude or longitude value, we rescale to a floating point value. With this technique, if two geolocations in the original graph have a k-bit matching prefix in their scaled and shifted values, then the resulting values will also have the same length prefix-match.

2.2.3 Tier 3 Obfuscation. Tier 3 represents the highest level of obfuscation, but also the lowest level of utility with respect to any location details retained in the map. The method employed at tier 3 is to randomly generate nodes and edges based on the latitude/longitude of existing nodes and lengths of existing edges. The tier 3 obfuscator, requires a user-provided probability ("goal probability") parameter that is used to determine whether any node or edge in the original map is included in the obfuscated map. A lower goal probability value provides stronger resistance against structural deobfuscation attacks, but reduces the utility of the resulting obfuscated graphs.

Using an input map and goal probability, we compute normal distributions of node geolocations, node degrees and edge lengths. We then randomly remove nodes and their incident edges, taking into consideration the original node degrees such that nodes with higher degrees will be less likely to be removed. As a result, "hubs" in the original map will be largely preserved in the obfuscated map. Next, we randomly remove edges to attempt to satisfy the goal probability. Then, we add new nodes and edges by sampling values from the computed normal distributions. Finally, we add edges to any singleton nodes to ensure a fully connected graph. We then run a tier 1 or tier 2 obfuscation algorithm on the randomized graph. The resulting map thus includes the connectivity and distributional characteristics from the original physical map, but no other details. An obfuscated map resulting from tier 3 is thus the most difficult to reverse engineer, and has utility similar to graphs produced from synthetic network connectivity generators such as BRITE [44] or Orbis [48].

2.3 Implementation in Bokeh

The *Geocentric Fuzzing* algorithm was implemented in Python (about 50 lines of code) and the rest of the algorithms described above were implemented in C# (about 500 lines of code). The randomization algorithm used in tier 3 can be run in combination with any of the other algorithms implemented in C#. We only show results from combining randomization with *Precision Truncation*. As input to Bokeh,

the goal probability for tier 3 or parameters to affect precision truncation or accuracy fuzzing can be provided.

Bokeh takes two types of input files: shapefiles (.shp format) and line-of-sight (LOS) graphs (.csv format). It outputs graphs in the same format used for input. Without any form of parallelization and running on a modestly provisioned PC, Bokeh's cryptographic obfuscation takes less than 20 minutes to obfuscate large shapefiles with hundreds of nodes and links (*e.g.*, the map of CenturyLink's communication infrastructure). All other algorithms take fewer than 8 minutes to obfuscate large shapefiles. We plan to improve the efficiency of Bokeh *e.g.*, through parallelization as part of our ongoing work.

3 APPLYING BOKEH TO ISP MAPS

In this section, we demonstrate Bokeh through a case study in which we apply it to several Internet infrastructure maps. We start by describing the base maps and then highlight the effects of each tier in removing details after applying Bokeh.

3.1 Data

To demonstrate the efficacy of Bokeh, we use service providers from the Internet Atlas project [16]. Internet Atlas includes over 1.5k maps of physical Internet infrastructure maps collected from public sources via web search. Each map includes the detailed geography of nodes (*e.g.*, POPs, data centers, co-location centers, etc.) and fiber conduits/links. Links details vary from LOS connectivity—represented as adjacency matrix in the repository—to detailed fiber conduit-level information stored in native GIS formats (*e.g.*, .shp). We consider the **fiber map** for CenturyLink (together with Level3, which was acquired by CenturyLink) with detailed infrastructure/conduit information, and **line-of-sight (LOS) maps** for Layer 42 (now Wave) and Aurora Fiber.

3.2 Results

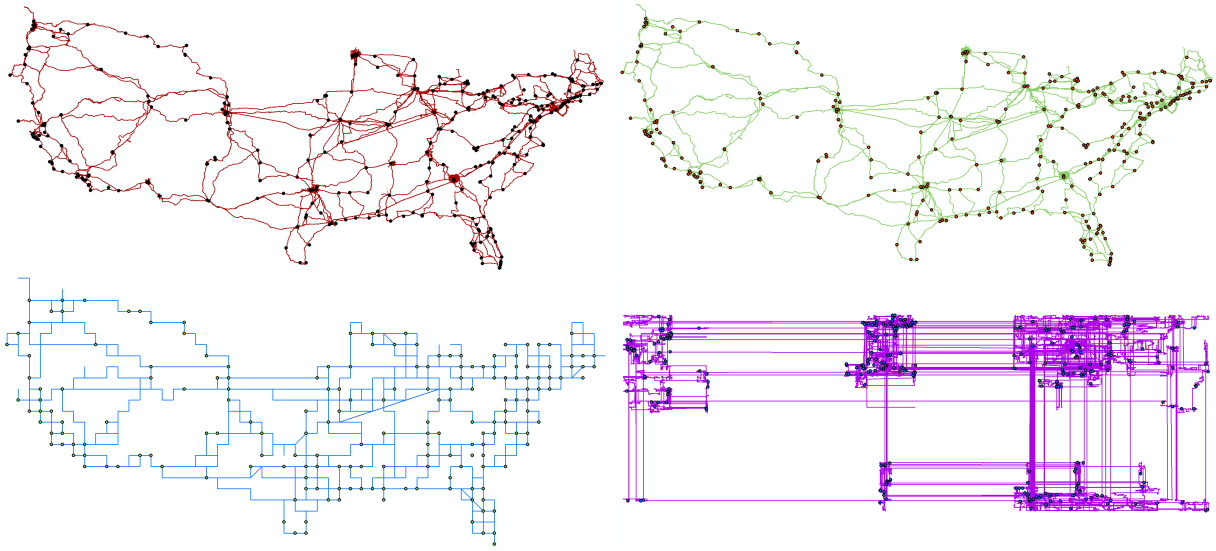
The goals of our assessment are to illustrate: *(i)* how Bokeh can remove or blur geographic details of Internet infrastructure maps across the three obfuscation tiers, and *(ii)* the tradeoffs in obfuscation versus utility for research. We use the following configuration settings for the results below: accuracy fuzzing uses p value of -2, precision truncation uses d value of 0, and the tier 3 algorithm uses a goal node probability value of 0.8 and goal edge probability value of 0.9. All other algorithms use randomized seeds.

Excising sensitive details using Bokeh. Figure 1a shows the results of Bokeh applied to the CenturyLink/Level3 map, at each of the three obfuscation tiers.[3] Figure 1a (upper left) shows the base map of CenturyLink network with detailed fiber and node information in red and black, respectively. The application of geocentric fuzzing, precision truncation, and cryptographic fuzzing are also shown in the figure. We observe the following. *(i)* Although geocentric fuzzing modifies the accuracy of geographic details, the precision is preserved as evident from the resulting map (Figure 1a, upper right). Since the outcome is qualitatively similar to the base map, except for the shift in accuracy of the geographic coordinates, we do not show the results of geocentric fuzzing for the other two network maps. *(ii)* Due to modifications to the precision of geographic coordinates, the effect of precision truncation on links with detailed fiber information is somewhat different from geocentric fuzzing despite the fact that both techniques belong to the same obfuscation tier. The relationship mentioned above to k-anonymity is apparent in the map due to some nodes and links "collapsing" to the same obfuscated location. *(iii)* The cryptographic fuzzing method ensures that the general structure of the network is preserved—the nodes on the west coast cluster together on the west region; similarly we observe clusters for mid-west, north-east and south-east. Despite preserving the general structure of the network map, this method ensures that an adversary would not be able to recover the original latitude/longitude values from the coordinates. *(iv)* Given the presence of diverse features (*i.e.*, lines and points) in maps such as those for CenturyLink/Level3, our tier 3 randomizer obfuscation cannot yet be applied to shapefiles. We are investigating this in on-going work.

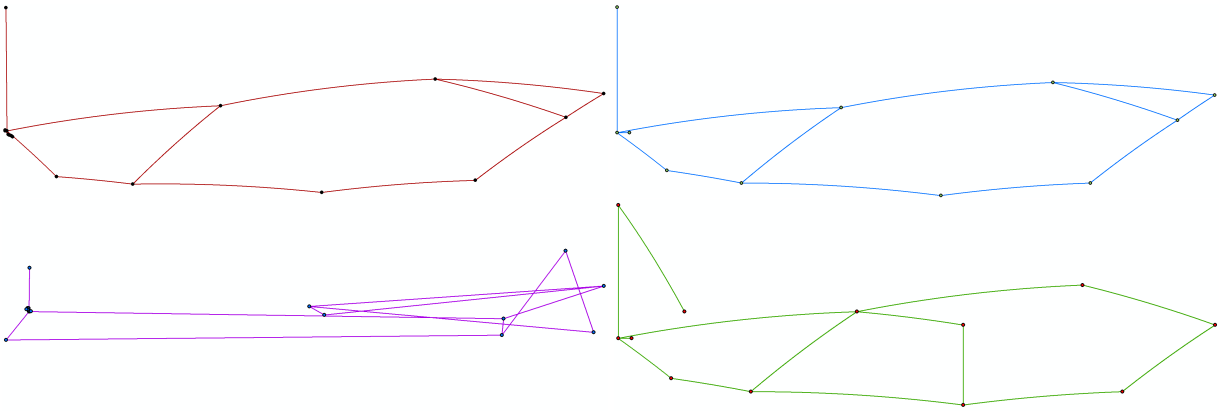
Figures 1b and 1c show the results of applying Bokeh to LOS maps of Layer 42 (now Wave) and Aurora Fiber. We observe the following. *(i)* The structure of maps resulting from tier 1, in comparison to the corresponding base maps, are qualitatively similar. *(ii)* Randomized (tier 3) fuzzing is applicable to the LOS maps and offers the highest level of obfuscation. *(iii)* Although the individual locations are obfuscated using cryptographic fuzzing, locations in the top-right region of Minnesota are grouped together due to a k-bit prefix match of their latitude/longitude after obfuscation before scaling. In addition, we note that the maps resulting from accuracy fuzzing are visually/qualitatively similar to geocentric fuzzing. Moreover, the TCPdpriv algorithm preserves intra-neighborhood connections quite well. However, the inter-neighborhood connections hinder visualization of maps.

Utility. To complement the visualizations shown in Figure 1 , we next show how Bokeh preserves the utility of the infrastructure maps. Figure 2 shows the CDFs of the geographic displacement of nodes and fiber conduits (in meters) produced by the tier 1 obfuscation methods with the aforementioned parameterization on CenturyLink (now Level3), Layer 42 (now Wave), and Aurora fiber maps. We observe that 90% of the geolocation values move a maximum distance of 1.5 km with the application of geocentric

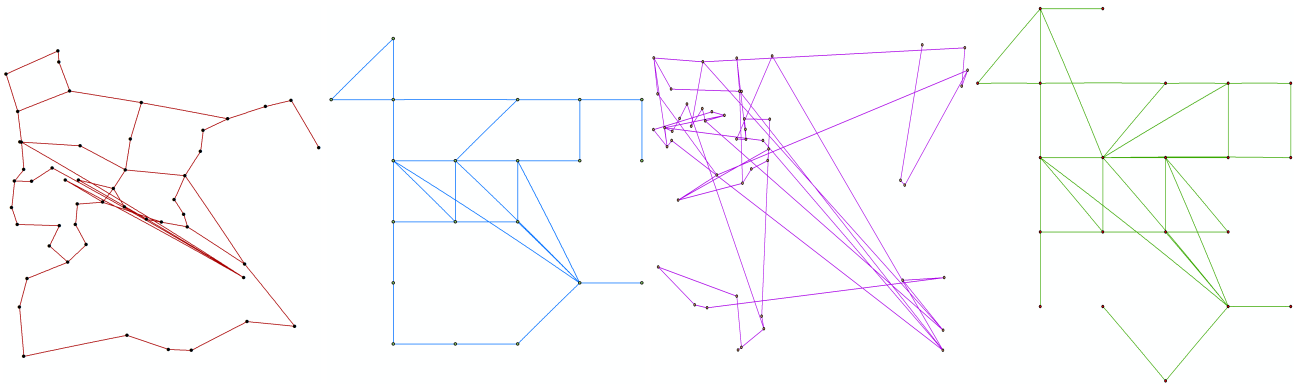
[3] Results of applying Bokeh to other maps from Internet Atlas are qualitatively similar and omitted for brevity.



(a) Basemap - red / black, geocentric fuzzing - green / red and precision truncation - blue / green (tier 1), and cryptographic - violet / blue (tier 2) fuzzing techniques of Bokeh applied on CenturyLink, which are large world-wide networks.



(b) Basemap - red / black, precision truncation - blue / green (tier 1), cryptographic fuzzing - violet / blue (tier 2), and randomized fuzzing - green / red (tier 3) of Layer 42 (now Wave), a regional network.



(c) Basemap - red / black, precision truncation - blue / green (tier 1), cryptographic fuzzing - violet / blue (tier 2), and randomized fuzzing - green / red (tier 3) of Aurora fiber, a metro network.

Figure 1: Tiers of obfuscation produced by Bokeh for a collection of physical infrastructure maps from Internet Atlas repository.

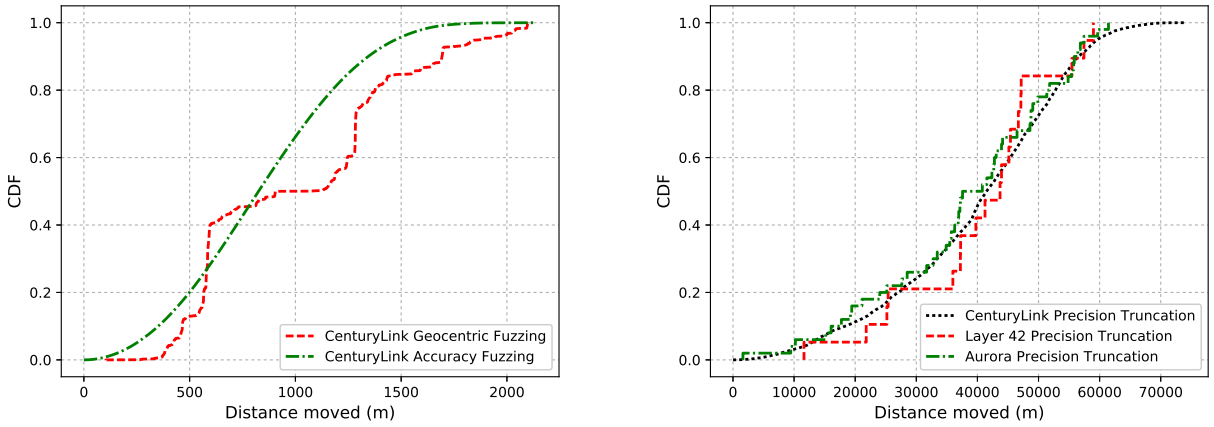


Figure 2: CDF showing the geographic displacement of nodes and links (in meters) by tier 1 obfuscation methods on CenturyLink, Layer 42 (now Wave), and Aurora fiber maps.

or accuracy fuzzing. Additionally, we observe that 90% of geolocation values move a maximum distance of 60 km with the application of precision truncation algorithm on all the network maps. These results indicate that the level of obfuscation is moderate *without* compromising the essential utility and geographical representation of infrastructure maps.

For tier 3 using LOS maps, the distributional properties of maps are preserved in Bokeh while offering the highest level of obfuscation. For high values of goal node and edge probabilities, the mean and standard deviations of latitude and longitude do not change by more than 10%. Additionally, the average node degree does not change by more than 10%. Average edge length does not change by more than 20%. For very low values of goal node and edge probabilities, all the metrics can change up to 50%. These metric values are consistent across all the LOS maps.

4 DISCUSSION

Broader Applicability. While our case study highlights obfuscation of maps of Internet infrastructure, our methodology and the Bokeh tool are general and can be applied to any maps that may be considered sensitive either in terms of graph or location characteristics. Examples include maps of *(i)* power grids including generation locations and transmission lines, *(ii)* hydrographic (water supply) delivery systems and other local services, and *(iii)* user location or route data that has been a topic in the popular press recently (*e.g.,* [31]). In each of these examples and others, it is common for maps to be represented as GIS shapefiles, which can be obfuscated via Bokeh.

Threat Analysis. The primary threat to obfuscated shapefiles is deobfuscation *i.e.,* determining the true locations of nodes and links. As noted in Section 1, we have yet to conduct a formal analysis of the resilience of our methodology to attackers. While our tiered methodology is informed by prior studies *that include formal analysis*, and offers a range of map obfuscation options that present increasing challenges to an adversary, we expect that a determined attacker with sufficient understanding of shapefiles and physical networks may be able to deobfuscate a map or set of maps that are obfuscated with minimal alteration from the original maps [13]. A comprehensive formal analysis is the subject of on-going work.

In the meantime, we posit the following framework for considering deobfuscation threats. Attackers can employ several different methods to recover detail in maps/shapefiles that is removed/altered during the obfuscation process. We note that shapefiles that are obfuscated via Bokeh do not have any artifacts of the original locations and connectivity. Thus, the primary attack method is to *infer* the original locations and connectivity.

We assume a strong attacker with access to a variety of data sources including *(i)* an ensemble of obfuscated maps/shapefiles of physical infrastructure, *(ii)* unobfuscated infrastructure maps/shapefiles found via search [16], *(iii)* maps/shapefiles or images *e.g.,* from Google maps of other infrastructures such as road, rail, co-location centers, etc. (in the case of communication infrastructure, nodes/links are often deployed along these rights of way [15]), or *(iv)* by other types of information or measurements that confine the geographic scope of infrastructures to specific locations/cities. The attacker can potentially infer locations or likely locations of links/nodes by using location averaging or by pinning certain key features of maps (such as buildings or facilities that

advertise their locations on line) and then inferring likely paths for links (*e.g.,* via satellite images).

Defending against such a strong attacker implies a conservative approach for obfuscation. Shapefile owners should assume determined attackers will get access to obfuscated maps. Thus, the tier and the parameters for obfuscation within that tier should be selected in a way that balances the utility of the resulting maps for envisioned research studies with the potential risk for deobfuscation. We are considering obfuscation metrics in on-going work (see below), which will enable assessment of both utility and difficulty in deobfuscation.

Quantifying Obfuscation. Beyond Figure 2, we do not provide a quantitative analysis of obfuscation in terms of specific *accuracy metrics*. This too is a subject of on-going work. We consider accuracy in terms of deviation from ground truth for both graph and location details. There is substantial prior work on metrics for comparisons of network graph properties (*e.g.,* [6]). Similarly, the issue of location accuracy metrics has been the subject of recent studies (*e.g.,* [3]).

5 RELATED WORK

Anonymization in social networks and mobile apps. Application of data anonymization techniques to social network graphs has been a significant focus in recent years. Liu *et al.* [39] propose k-degree anonymity, which prevents node degree-based attacks by ensuring that for each node there are $k-1$ other nodes with the same degree; a modified approach by Lu *et al.* [40] was shown to be more effective for anonymization of large real-world graphs. Other graph-based k-anonymization techniques include k-neighborhood [59, 64], k-automorphism [65], k-isomorphism [11], k-symmetry [61] and evolutionary k-degree anonymization [9]. Bhagat *et al.* [5] propose class-based anonymization algorithms which partition entities into classes. They also release edge connectivity information between the classes. Nguyen *et al.* [47] apply Maximum Variance (MV) approach for social graph anonymization based on edge uncertainty semantics. Various types of de-anonymization attacks and algorithms that are robust to such attacks have been the focus of many efforts [2, 4, 29, 33, 34, 38, 46]. Work by Hasan *et al.* [28] on anonymization of user trajectories for location-based service applications bears resemblance to our work. They adopt a bounded perturbation algorithm and add noise to geolocations taking into consideration the validity of the new geolocation generated. Our tier-1 algorithms apply similar techniques to obfuscate geographic points in maps. Finally, there have been many studies on location data privacy in smartphones apps (*e.g.,* [23, 26]). These studies are related to ours in that they are concerned with limiting exposure of lat-lon location details. However, these studies by definition are focused on individual devices that run diverse applications, whereas our work is focused on obfuscation of shapefiles (.shp) and LOS (.csv) maps.

Internet data privacy. Anonymizing personally identifiable information (*e.g.* IP addresses) in packet traces captured in the live Internet is essential to ensure the privacy of senders and receivers. TCPdpriv [24] performs prefix-preserving pseudonymization, retaining the prefix relationship between IP addresses after processing. Xu *et al.* [62, 63] addressed limitations of applying TCPdpriv in a large-scale distributed setting by proposing a cryptographic approach. Our tier-2 algorithms leverage both of these prior techniques to obfuscate geolocations in physical network maps. Other prior works in IP address obfuscation include ipsumdump [18], tcpurify [17], tcpanon [21], tcpmkpub [49], and [50].

Internet mapping. Analyzing and generating maps and graphs that represent Internet communication infrastructure has been the subject of many prior studies. Notable efforts include Internet Atlas [15, 16] and Topology Zoo [35]. Each map in these repositories consists of potentially sensitive information including geographic location of nodes (*e.g.*, POPs, data centers, co-locations) and the links that interconnect these nodes (*e.g.*, fiber-optic links). Our techniques will remove or blur sensitive details from such maps. In addition to these efforts, various studies have focused on inferring router-level network topologies [8, 42, 54, 55]. Prior efforts have also included modeling router-level topologies [22, 37, 43, 44, 58, 60]. Apart from obfuscating the maps from these efforts, Bokeh could also be applied to router geolocation information [25].

6 SUMMARY

Maps of physical infrastructure that include details of deployment locations and connectivity are important in a variety of research areas including assessing risks (*e.g.,* to severe storms) and opportunities to improve robustness, performance and security. They are also valuable for growth planning, inventory management, recommendation systems, e-marketing and other locality-based services. However, maps like these can be used for unwanted or malicious purposes and can therefore be considered a risk if shared openly.

In this paper, we address the problem of obfuscating maps of physical infrastructure. Our motivation is to develop a capability for obfuscation that can lower risks and thus encourage providers to make more infrastructure maps available for research. A key objective is to develop a method that includes a range of obfuscation capabilities that enable risk to be addressed by map providers while preserving the utility of the maps for research. We describe a three tiered methodology that affects progressively higher levels of map

obfuscation from simple location fuzzing through transformations that only preserves distributional properties of a network.

We implement our methodology in a tool called Bokeh. Bokeh accepts standard representations of network infrastructure maps (shapefiles or csv) as input and produces maps with obfuscated locations and/or connectivity characteristics. We demonstrate Bokeh in a case study on a selection of maps of Internet Service Provider networks and show how the resulting maps are increasingly obfuscated from the original. We argue that Bokeh is useful in a broad context for obfuscation of maps of any kind of infrastructure (*i.e.,* any map represented in supported input format). The Bokeh source code and a set of 42 obfuscated maps of Internet Service Provider networks are available to the research community at https://github.com/yugaligullapalli/Bokeh and https://www.impactcybertrust.org respectively.

ACKNOWLEDGEMENTS

This work is supported by NSF CNS-1703592, NSF CNS-1814537, DHS BAA 11-01, AFRL FA8750-12-2-0328. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, DHS, AFRL or the U.S. Government.

REFERENCES

[1] [n.d.]. More Insights On Alleged DDoS Attack Against Liberia Using Mirai Botnet. https://thehackernews.com/2016/11/ddos-attack-mirai-liberia.html.

[2] Charu C Aggarwal, Yao Li, and S Yu Philip. 2011. On the hardness of graph anonymization. In *Data Mining (ICDM)*. IEEE.

[3] Hidaytet Aksu, Demet Aksoy, and Ibrahim Korpeogluy. 2013. A Study of Localization Metrics: Evaluation of Position Errors in Wireless Sensor Networks. *Computer Networks* 55, 15 (2013).

[4] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. 2007. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*. ACM.

[5] Smriti Bhagat, Graham Cormode, Balachander Krishnamurthy, and Divesh Srivastava. 2009. Class-based graph anonymization for social network data. *Proceedings of the VLDB Endowment* 2, 1 (2009).

[6] Alireza Bigdeli, Ali Tizghadam, and Alberto Leon-Garcia. 2009. Comparison of Network Criticality, Algebraic Connectivity, and other Graph Metrics. In *Proceedings of the 1st Annual Workshop on Simplifying Complex Network for Practitioners*. ACM.

[7] Ilker Nadi Bozkurt, Anthony Aguirre, Balakrishnan Chandrasekaran, P. Brighten Godfrey, Gregory Laughlin, Bruce Maggs, and Ankit Singla. 2017. Why Is the Internet so Slow?!. In *Proceedings of the Passive and Active Measurement Conference*.

[8] CAIDA. 2018. Archipelago (Ark) Measurement Infrastructure. http://www.caida.org/projects/ark/.

[9] Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. 2013. Evolutionary algorithm for graph anonymization. *arXiv preprint arXiv:1310.0229* (2013).

[10] Joseph Chabarek, Joel Sommers, Paul Barford, Cristian Estan, David Tsiang, and Steve Wright. 2008. Power awareness in network design and routing. In *INFOCOM. The 27th Conference on Computer Communications*. IEEE.

[11] James Cheng, Ada Wai-chee Fu, and Jia Liu. 2010. K-isomorphism: privacy preserving network publication against structural attacks. In *Proceedings of the ACM SIGMOD International Conference on Management of data*. ACM.

[12] Jonathan Chew. 2015. Somebody is Cutting Internet Cables, Causing Massive Outages. *Fortune Magazine* (2015).

[13] Scott Coull, Charles Wright, Fabian Monrose, Michael Collins, and Michael Reiter. 2007. Playing DevilÃŢs Advocate: Inferring Sensitive Information from Anonymized Network Traces. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*. Internet Society.

[14] Ramakrishnan Durairajan and Paul Barford. 2017. A Techno-Economic Framework for Broadband Deployment in Underserved Areas. *ACM SIGCOMM Computer Communication Review* 47, 2 (2017).

[15] Ramakrishnan Durairajan, Paul Barford, Joel Sommers, and Walter Willinger. 2015. InterTubes: A study of the US long-haul fiber-optic infrastructure. In *ACM SIGCOMM Computer Communication Review*, Vol. 45. ACM.

[16] Ramakrishnan Durairajan, Subhadip Ghosh, Xin Tang, Paul Barford, and Brian Eriksson. 2013. Internet Atlas: a Geographic Database of the Internet. In *Proceedings of the 5th ACM workshop on HotPlanet*. ACM.

[17] E. Blanton. 2018. Tcpurify. http://brewformulas.org/Tcpurify.

[18] E. Kohler. 2018. IPSUMDUMP. http://read.seas.harvard.edu/~kohler/ipsumdump/.

[19] Brian Eriksson and Mark Crovella. 2013. Understanding geolocation accuracy using network geometry. In *INFOCOM*. IEEE.

[20] ESRI. 2018. ESRI ArcGIS Shapefiles. https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm.

[21] F. Gringoli. 2018. tcpanon. http://netweb.ing.unibs.it/~ntw/tools/tcpanon/.

[22] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, Vol. 29. ACM.

[23] K. Fawaz and K. Shin. 2014. Location Privacy Protection for Smartphone Users. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

[24] G. Minshall. 2018. TCPDPRIV. http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html.

[25] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. 2017. A Look at Router Geolocation in Public and Commercial Databases. In *Proceedings of the ACM Internet Measurement Conference*. ACM.

[26] S. Guha, M. Jain, and V. Padmanabhan. 2012. Koi: a location-privacy platform for smartphone apps. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. ACM.

[27] John W Hager, James F Behensky, and Brad W Drew. 1989. *The Universal Grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS). Edition 1.* Technical Report. DEFENSE MAPPING AGENCY HYDROGRAPHIC/TOPOGRAPHIC CENTER WASHINGTON DC.

[28] ASM Hasan, Qiang Qu, Chengming Li, Lifei Chen, and Qingshan Jiang. 2018. An Effective Privacy Architecture to Preserve User Trajectories in Reward-Based LBS Applications. *ISPRS International Journal of Geo-Information* 7, 2 (2018).

[29] Michael Hay, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. 2008. Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment* 1, 1 (2008).

[30] Brandon Heller, Rob Sherwood, and Nick McKeown. 2012. The Controller Placement Problem. In *Proceedings of the First Workshop on Hot Topics in Software Defined Networks*. ACM.

[31] Alex Hern. 2018. Fitness Tracking App Strava Gives Away Locations of Secret US Army Bases. *The Guardian* (2018).

[32] Internet2. 2014. Network Flow Data Privacy Policy. https://www.internet2.edu/policies/network-flow-data-privacy-policy/.

[33] Shouling Ji, Weiqing Li, Prateek Mittal, Xin Hu, and Raheem A Beyah. 2015. SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization.. In *USENIX Security Symposium*.

[34] Shouling Ji, Weiqing Li, Mudhakar Srivatsa, and Raheem Beyah. 2014. Structural data de-anonymization: Quantification, practice, and implications. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM.

[35] Simon Knight, Hung X Nguyen, Nick Falkner, Rhys Bowden, and Matthew Roughan. 2011. The Internet Topology Zoo. *IEEE Journal on Selected Areas in Communications* 29, 9 (2011).

[36] Sumeeet Kumar and Kathleen Carley. 2017. Simulating DDOS attacks on the us fiber-optics internet infrastructure. In *Proceedings of the Winter Simulation Conference*. IEEE.

[37] Lun Li, David Alderson, Walter Willinger, and John Doyle. 2004. A first-principles approach to understanding the Internet's router-level topology. In *ACM SIGCOMM Computer Communication Review*, Vol. 34. ACM.

[38] Yidong Li and Hong Shen. 2010. Anonymizing graphs against weight-based attacks. In *Data Mining Workshops (ICDMW)*. IEEE.

[39] Kun Liu and Evimaria Terzi. 2008. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD international conference on Management of data*. ACM.

[40] Xuesong Lu, Yi Song, and Stéphane Bressan. 2012. Fast identity anonymization on graphs. In *International Conference on Database and Expert Systems Applications*. Springer.

[41] Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, David Clark, et al. 2016. Bdrmap: Inference of borders between IP networks. In *Proceedings of the 2016 Internet Measurement Conference*. ACM.

[42] Harsha V Madhyastha, Tomas Isdal, Michael Piatek, Colin Dixon, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. 2006. iPlane: An information plane for distributed services. In *Proceedings of the 7th symposium on Operating systems design and implementation*. USENIX Association.

[43] Priya Mahadevan, Calvin Hubble, Dmitri Krioukov, Bradley Huffaker, and Amin Vahdat. 2007. Orbis: rescaling degree correlations to generate annotated Internet topologies. In *ACM SIGCOMM Computer Communication Review*, Vol. 37. ACM.

[44] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. 2001. BRITE: An approach to universal topology generation. In *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2001. Proceedings. Ninth International Symposium on*. IEEE.

[45] Roland Meier, Petar Tsankov, Vincent Lenders, Laurent Vanbever, and Martin Vechev. 2018. NetHide: secure and practical network topology obfuscation. In *Proceedings of the 27th USENIX Security Symposium*. 693–709.

[46] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Security and Privacy*. IEEE.

[47] Hiep H Nguyen, Abdessamad Imine, and Michael Rusinowitch. 2014. A maximum variance approach for graph anonymization. In *International Symposium on Foundations and Practice of Security*. Springer.

[48] P. Mahadevan. 2018. Analyzing and Generating Network Topologies with Orbis. http://www.sysnet.ucsd.edu/~pmahadevan/topo_research/topo.html.

[49] Ruoming Pang, Mark Allman, Vern Paxson, and Jason Lee. 2006. The devil and packet trace anonymization. *ACM SIGCOMM Computer Communication Review* 36, 1 (2006).

[50] Ruoming Pang and Vern Paxson. 2003. A high-level programming environment for packet trace anonymization and transformation. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications*. ACM.

[51] PROJ contributors. 2018. *PROJ coordinate transformation software library*. Open Source Geospatial Foundation. http://proj4.org/

[52] R. Kulkarni. 2018. A Dissertation So Good It Might Be Classified. https://www.wired.com/2004/01/a-dissertation-so-good-it-might-be-classified/.

[53] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Technical Report. Technical report, SRI International.

[54] Yuval Shavitt and Eran Shir. 2005. DIMES: Let the Internet measure itself. *ACM SIGCOMM Computer Communication Review* 35, 5 (2005).

[55] Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson. 2004. Measuring ISP topologies with Rocketfuel. *IEEE/ACM Transactions on Networking (ToN)* 12, 1 (2004).

[56] Volker Stocker, Georgios Smaragdakis, William Lehr, and Steven Bauer. 2016. Content may be King, but (Peering) Location matters: A Progress Report on the Evolution of Content Delivery in the Internet. In *Proceedings of the 27th European Regional Conference*. International Telecommunications Society (ITS).

[57] Henry Stommel. 2017. *Lost islands: The story of islands that have vanished from nautical charts*. Courier Dover Publications.

[58] Hongsuda Tangmunarunkit, Ramesh Govindan, Sugih Jamin, Scott Shenker, and Walter Willinger. 2002. Network topology generators: Degree-based vs. structural. In *ACM SIGCOMM Computer Communication Review*, Vol. 32. ACM.

[59] BK Tripathy and GK Panda. 2010. A new approach to manage security against neighborhood attacks in social networks. In *Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

[60] Bernard M Waxman. 1988. Routing of multipoint connections. *IEEE journal on selected areas in communications* 6, 9 (1988).

[61] Wentao Wu, Yanghua Xiao, Wei Wang, Zhenying He, and Zhihui Wang. 2010. K-symmetry model for identity anonymization in social networks. In *Proceedings of the 13th international conference on extending database technology*. ACM.

[62] Jun Xu, Jinliang Fan, Mostafa Ammar, and Sue B Moon. 2001. On the design and performance of prefix-preserving IP traffic trace anonymization. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. ACM.

[63] Jun Xu, Jinliang Fan, Mostafa H Ammar, and Sue B Moon. 2002. Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. In *Network Protocols*. IEEE.

[64] Bin Zhou and Jian Pei. 2008. Preserving privacy in social networks against neighborhood attacks. In *Data Engineering. ICDE*. IEEE.

[65] Lei Zou, Lei Chen, and M Tamer Özsu. 2009. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment* 2, 1 (2009).