Network Report: A Structured Description for Network Datasets

Xinyi Zheng
Carnegie Mellon University

Ryan A. Rossi
Adobe Research

Nesreen Ahmed
Intel Labs

Dominik Moritz
Carnegie Mellon University

ABSTRACT

The rapid development of network science and technologies depends on shareable datasets. Currently, there is no standard practice for reporting and sharing network datasets. Some network dataset providers only share links, while others provide some contexts or basic statistics. As a result, critical information may be unintentionally dropped, and network dataset consumers may misunderstand or overlook critical aspects. Inappropriately using a network dataset can lead to severe consequences (e.g., discrimination) especially when machine learning models on networks are deployed in high-stake domains. Challenges arise as networks are often used across different domains (e.g., network science, physics, etc) and have complex structures. To facilitate the communication between network dataset providers and consumers, we propose network report. A network report is a structured description that summarizes and contextualizes a network dataset. Network report extends the idea of dataset reports (e.g., Datasheets for Datasets) from prior work with network-specific descriptions of the non-i.i.d. nature, demographic information, network characteristics, etc. We hope network reports encourage transparency and accountability in network research and development across different fields.

ACM Reference Format:
Xinyi Zheng, Ryan A. Rossi, Nesreen Ahmed, and Dominik Moritz. 2022. Network Report: A Structured Description for Network Datasets. In Proceedings of ACM Conference (Conference’17). ACM, New York, NY, USA, 11 pages. doi:https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Scientists use networks as representation to understand and model complex systems including social, biological, neural, and technological systems [41]. Networks are also at the heart of industrial applications, empowering products from Google [45] to Facebook, CISCO, and Twitter. Recently, machine (deep) learning on graphs and networks has received increasing attention in major conferences, driving real-world applications such as drug discovery [61], recommender systems [60] and protein structure prediction [25].

Sharing and using standard datasets accelerate advancements in science and technology [2]. Through our random investigation on over 10 papers presenting new network datasets (including their associated websites), we found that they describe network datasets in only 1-2 short paragraphs listing the genre and basic statistics. As different network dataset providers make their own effort in documenting the datasets, important information is often unintentionally omitted. Various data quality issues in common network benchmarks have been pointed out recently [22, 53].

In natural language processing (NLP), researchers have found that critical information such as ‘what language is included’ is often omitted [3]. Proposals have been made and adopted to standardize how to document NLP datasets [4]. To address the gap in documenting datasets for machine learning (ML) more broadly, Gebru et al. [16] proposed Datasheets for Datasets, which quickly became the norm for data reporting in industry and academia.

However, networks are ubiquitous and have unique properties that warrant an extension of these formats. First, networks are created and researched across different areas and domains (e.g., computer science, physics, biology, etc). Assumptions and practices are not always shared across these communities, therefore increasing the need for structured communication. Take for example a social network dataset. Social science researchers build the social network through surveys, interviews, etc, and conduct social science research. Network science researchers study the community structure. ML researchers study how to predict links. High performance computing researchers study how to compute social networks efficiently. Thus reporting network datasets should generalize to different areas and domains. Second, networks have complex structures and (often) contain thousands (millions) of nodes, so it is hard to communicate what a network ‘looks like’. Lists and tables are often easier to illustrate. Often one can show a representative instances of e.g., a computer vision dataset (images with object annotations as shown in Figure 2). In contrast, for networks, visualizing an instance (entity/node) and its neighbourhood does not reveal the structure of the whole network (sometimes even visualizing the neighbourhood could be hard). Another way to communicate what a network ‘looks like’ is to visualize the whole network. Unfortunately, visualizing large networks has proven to be difficult (as shown in Figure 1) [30]. Thus, basic statistics (e.g., degree) over networks are used as a tool to communicate network characteristics. Moreover, the relational nature of networks

Figure 1: Spring positioned visualization for a Watts Strogatz Graph [43] with 1000 nodes and 3204 edges

Figure 2: Image examples with instance segmentation from Coco dataset on its official website [33]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference’17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-xxxx-xxxx-x/YY/MM... $15.00
https://doi.org/XXXXXXX.XXXXXXX
make data collection and preparation more complex. There are data preparation operations unique to networks that are not covered by previous proposals for reporting datasets. For example, dataset providers can transform a network dataset based on its structure (e.g., extracting k-cores). Some data preparation operations that may be less emphasized in tabular data preparation could have unintended and complex effects on the network structure due to the non-i.i.d. nature of networks (i.e., nodes/samples are not observed i.i.d.). For instance, randomly sampling a scale-free network (i.e., degree distributions that roughly follow a power law) could result in a sub-network that is not scale-free [51]. These statistical pitfalls should be documented by network providers.

Current practice in reporting network datasets could result in inappropriate documentation leading to information gaps between network dataset providers and consumers. Thus, (unintentional) dataset misuse and misunderstanding could increase, adversely affecting science and society. Deployment of applications evaluated or trained with inappropriate networks could have serious consequences such as discrimination against minority groups [50].

To address the information gap between network dataset providers and consumers, we design network reports as practice for documenting network datasets. Our development was driven by formative interviews with 12 network practitioners (including faculty, graduate students, research scientists, and engineers in industry). We sought to understand what they would like to see when they encounter a new network dataset. A network report is a structured description following a particular template that summarizes and contextualizes a network dataset. To handle the unique challenges of reporting networks, we advocate for standard measures (e.g., degree) of the network structure, demographics information (if the dataset is related to people), potential bias that already exists or was introduced during data processing, and potential or known errors in the dataset. For network dataset providers, we anticipate that network report could relieve the burden of thinking about what to include and how to present the dataset intuitively. We anticipate that network dataset consumers can assess whether the dataset is appropriate for their task, compare different datasets, and understand limitations. We hope that network reports prevent unintentional misuse of network datasets and mitigate bias in high-stakes applications. While we think that the proposed format of network report covers the most important properties of networks, we see it as a living format that the community continues to refine.

2 RELATED WORK
A growing body of literature designs and evaluates structured documents or checklists for datasets, models, and services. One of the most prominent works is ‘Datasheets for Datasets’ [16]. The authors proposed to accompany documentation with datasets to encourage transparency and accountability for data used by the ML community. The framework advocates for systematically documenting datasets’ purpose, composition, collection process, preprocessing, uses, distribution, and maintenance. A similar proposal is ‘Data statements for NLP’, which focuses on NLP datasets and addresses ethics, exclusion, and bias issues in NLP datasets and systems [4]. ‘Dataset nutrition labels’ describes the concept of nutrition label to assess the quality and robustness of a dataset before it is used in model development [21, 58]. Recent studies have also investigated various aspects of documentation in data science lifecycle such as data readiness [5], industrial data practices for computer vision datasets [37], academic dataset usage [17], accountability for ML datasets [23] etc. Beyond datasets, ‘model cards’ were proposed to accompany trained ML models including information like intended use cases, evaluation data, performance measures etc [39]. The goal of ’model cards’ is to improve transparency between stakeholders of ML models. ‘FactSheets’ and ‘AI fairness checklists’ have a similar motivation, but were targeted at documenting AI services and systems [20, 35].

Complementing existing work, we primarily focus on networks, and develop a structured way to communicate network datasets. Our work is functionally similar to ‘Datasheets’ and ‘Data statements’ in the context of ML, but we also consider challenges outside ML as networks are used and shared across different areas and domains.

3 PRELIMINARIES
We briefly discuss some terms and symbols in this section. “Graph” is more prevalent in the ML community, but “network” has been popular in the data mining and network science communities. We primarily use “network” in this paper, and use “graph” for specific terms (e.g., knowledge graph). The reason is that network analysis is generally concerned with the properties of real-world data, whereas graph theory is concerned with the theoretical properties of the mathematical graph abstraction.

A network G is represented as a tuple $G = (V, E, X^V, X^E)$, where $V$ is the set of nodes, $E$ is the set of edges, $X^V$ is the attribute matrix of nodes, $X^E$ is the attribute matrix of edges. Nodes $V$ represent entities, and edges $E$ represent relationships between entities. $X^V$ represents node attributes. $X^E$ represents edge attributes. Each row of $X^V$ (X^E) represents one node (edge), and each column of $X^V$ (X^E) represents one attribute. D denotes the degree matrix, where the degree of each node is on the diagonal. A, L denote the adjacency matrix, the Laplacian matrix of the network respectively. We will use this terminology throughout the rest of the paper.

4 DEVELOPMENT PROCESS
We started with formative interviews with 12 network practitioners including faculty, graduate students, research scientists, and engineers in industry. We asked interviewees at least: what kind of network data they would use; what kind of graph problems they are (have been) working on, and what action they would take when encountering a new network dataset.

Next, using our experience as network researchers, we analyzed potential use cases of networks in academic and industrial contexts. In academic contexts, the network report should provide readers enough relevant information to understand the academic work. Given an arbitrary network, the best algorithm for an analytical problem (e.g., community detection) depends on the characteristics (e.g., size, cohesion, spectrum etc.) of the network. For instance, community detection based on asymptotical Surprise [54] works well for many small communities, while Louvain community detection [8] based on modularity favours networks with few large communities. This is due to the no free-lunch theorem, which
states that any two optimization algorithms are equivalent when averaged across all possible problems (datasets) [24]. As network algorithms could favor certain types of networks as opposed to others, when reading scientific work, researchers need sufficient information of the network datasets to understand the performance of the proposed methods, and how the proposed methods can or cannot be expected to achieve similar results on other datasets. The aforementioned examples highlight the need to include network characteristics measures.

In industrial contexts, deployment of applications evaluated or trained with inappropriate networks could have serious consequences [50]. A recommender system trained and deployed with different demographics, inappropriate time window or shifted customer inclinations could result in bad user experience and revenue lost [28]. Recent studies revealed that ML models reproduce or amplify unwanted societal biases reflected in training data [10]. For example, knowledge graph embeddings have been shown to encode that men are more likely to be bankers, and women more likely to be homekeepers [14]. A question answering system empowered by the knowledge graph may discriminate female. Adding a standardized documentation can make application developers be aware of the potential consequences when using the data and infer what additional steps are necessary. These examples highlight the need to document population, bias and fairness (for applications) when sharing a network dataset.

We prototyped a template for network reports based on what we learned. We then “tested” the template by creating the network report for the EU email dataset [32]. While creating the network report, we noted problems like non-intuitive orders, lack of clarity, lack of coverage, and redundancies. We then refined our initial template and recreated the above network report. We incorporated feedback from other network practitioners with two cases (https://tinyurl.com/mwwunkae#case) and finally created the template present in the next section. Finally, we conducted an initial user study in section 7, and incorporated the feedback from participants in the final version presented here.

5 PROPOSED TEMPLATE

Figure 3 shows the skeleton of the proposed template. We include four sections in network report: (i) Curation Rationale contains the metadata and the contexts of the dataset; (ii) Dataset Collection, Preprocessing and Annotation contains the data preparation information; (iii) Uses contains the usage information; (iv) Network Statistics contains the statistical properties of the dataset. The proposed sections are not intended to be exhaustive. Additional details may include domain-specific information, legal information, etc. Network dataset providers are also encouraged to review and integrate information in datasheets [16]. Here we focus on network aspects of documenting a dataset.

5.1 Curation Rationale

This section helps network dataset consumers understand the basic metadata of the dataset, and sets the context for the remaining sections. After reading this section in the report, readers should have a good overview of the dataset.

Author(s) and reference of the network dataset: This acknowledges the authors who share the network. If the network dataset in an academic paper, it is encouraged to list the reference.

Purpose: Why was the network dataset created? Was there a specific usage in mind? Network dataset providers typically create the dataset primarily for a specific use case. This point could inform...
Which includes every node, and every edge between nodes. Unfortunately, this is not feasible under most circumstances. As a result, there is no strict order for these steps, thus affecting all downstream data processing choices. In some domains, nodes or/and edges do not naturally exist in the raw data, and network dataset providers construct networks from the raw data. For instance, some brain networks are constructed from multiple time series signals, where each time series represents the activity for one region in the brain [47]. However, in tabular data, the meaning of each instance (sample) does not change (e.g., an image collected from the web is always an image in the dataset). Dataset providers should describe the raw data.

The data collection mechanism is a significant source of error when making measurements over network structures. In biology, metabolic, protein, and regulatory networks are typically measured in a lab environment. Suffering from natural variation in biological systems and inconsistent measurement conditions, experiments usually don’t give the same results every time, meaning that any individual measurement may be an error [42]. Social networks collected from surveys or questionnaires could suffer from missing data, where participants may decline to answer some or all questions [6, 36]. The error or/and bias introduced in the data collection should be included in network reports, thus network dataset consumers could infer whether the error or/and bias would influence their tasks.

Network sampling. Was there any sampling process when collecting the network dataset? What sampling strategy has been chosen? What was the reason for the sampling strategy? Network sampling often inherently happens when the data is collected. For example, researchers use breadth first search to crawl online social networks automatically [38]. Many factors make it difficult to study networks in their entirety. Some networks are not completely visible to the public (e.g., Facebook), or can be accessed only through crawling (e.g., the Web). In other cases, the measurements required to observe the underlying networks are costly (e.g., experiments in biological network). Moreover, many networks are continuously changing (e.g., telecommunication interactions, user-product interactions) and can never recorded in their entirety. For these networks, providers should describe the collection time frames.

The non-i.i.d. nature of networks makes network sampling complex. For example, given a population network, randomly sampling a set of nodes versus randomly sampling a set of edges would result in very different network structures. Network sampling could introduce sampling bias to the network, which makes it hard to correctly estimate the parameters of the whole network. For instance, sampling in a breadth-first manner is biased as high-degree nodes have a higher chance of being sampled [29]. People who intend to study hard-to-reach communities should not use breadth-first sampled social networks.

With the detailed sampling information, network dataset consumers could examine whether the network dataset is biased towards particular properties and whether it is still suitable for their task, and what additional processing steps are needed. For instance, researchers have used Metropolis-Hasting random walk to induce an unbiased sample of Facebook users [18].
Network dataset providers can interchange them under different circumstances. If network dataset providers provide the raw data, the data preprocessing part in network report allows network dataset consumers to replicate the steps.

**Network construction**: What strategies were used to transform raw data to nodes and edges? What are the reasons for choosing these strategies? When the raw data does not naturally include nodes or/and edges, network construction is a crucial step to set up nodes or/and edges. Essentially, network construction is to set up \( V \) and \( E \) in the network. This step is unique to network datasets as tabular data does not have relations between instances. Network construction determines the semantics of nodes or/and edges, so it should be consistent with ‘node and edge semantics’ in ‘Curation Rationale’. Different network construction methods fundamentally change the network structure and the compatible tasks. For example, in brain networks, the edges could represent anatomical tracts or functional associations, and the semantics largely determines the neurobiological interpretation of network topology [47].

**Data cleaning**: How were nodes, edges, or networks (in a multi-network dataset) edited or augmented? Data cleaning refers to the removal, addition, or replacement of less accurate (or inaccurate) data values with more suitable, accurate, or representative values [19]. Research on data cleaning for networks has shown that inappropriate data cleaning strategies could lead to a high error rate in wireless sensor networks due to power exhaustion and environmental noise [11]. Thus, including data cleaning details in a network report is necessary for network dataset consumers to examine if the data quality is appropriate for the tasks and if they need to conduct additional data cleaning steps. Typical operations of data cleaning on networks include: removal of nodes, removal of edges, removal of attributes, node deduplication, etc. Data cleaning has a direct impact on data quality and data usage. Removal of specific attributes like timestamps could limit the applicability of the network dataset.

**Data filtering**: What nodes or/and edges have been filtered out? What was the reason for filtering? Data filtering refers to generating a subset of nodes or/and edges [19]. Data cleaning focuses on removing inaccurate values, while data filtering focuses on simplifying the contents of the data. Different from tabular data, where data filtering is done based on attribute value(s) (value filtering), in networks, data filtering could be based on the network structure (structural filtering). Value filtering is to simplify contents based on \( X^V \) and \( X^E \), while structural filtering is based on \( V \) and \( E \). Structural data filtering could be a potential source of bias in data splitting strategy and provide details on how the strategies were chosen. In addition, the notion of ‘transductive learning’ and ‘inductive learning’ brings additional ambiguity when evaluating ML models on graphs and networks. In the context of graph machine learning, ‘transductive learning’ refers to seeing the whole graph (including the test set but without the labels of the test set) beforehand, while ‘inductive learning’ assumes there are unseen nodes added to the graph during test time. In an inductive node classification task, researchers have found duplicated nodes (near 5%), edges and labels in training and testing sets [22] due to data splitting. This violates the inductive assumption as the duplicated nodes in test sets cannot be viewed as newly added. To avoid similar issues, we encourage network dataset providers to carefully examine the data splitting strategy and provide details on how the strategies are implemented. We also encourage network dataset consumers, especially ML researchers, to check if the data is split appropriately for their tasks.

**5.2.3 Instance Demographics**: In case the network dataset is related to people, what are the demographic characteristics of the instances? Possible demographic information include but not limited to age, gender, race/ethnicity, and socioeconomic status. Including instance demographics is important for ML applications in real-world deployments. Researchers found that models can reproduce or amplify unwanted societal biases reflected in training data [10]. In real-world deployments, biased data can result in unfair decisions and discrimination of minority groups, which can lead to severe
consequences in high-stakes domains. For instance, recommender systems that make predictions based on observed data can easily inherit bias that may already exist [59]. Knowledge graph embeddings have been shown to encode that men are more likely to be bankers, and women are more likely to be homekeepers [14]. To mitigate bias and improve fairness, sensitive attributes (e.g., age in a loan and debt network; gender in a job posting network [7, 9]) involving discrimination should be available to network dataset consumers.

5.2.4 Data annotation. Who annotated the data (e.g., crowds, domain experts)? How was the data annotated? Was there a shareable user interface? Were the labels noisy or clean? How are labels derived from multiple noisy labels? The description of data annotation process is detailed in this part. Including details at the data annotation process in network reports is useful for network dataset consumers who would like to use the labels.

5.2.5 Annotator Demographics. What are the demographic characteristics of the annotators? Possible demographic information include but not limited to age, gender, race/ethnicity, socioeconomic status, native language, and expert level. In social networks measured using surveys or questionnaire, subjectivity is a notable source of error when measuring network structure [42], especially when respondents give ambiguous answers and some interpretation is required to decide what those answers mean. Annotators’ native language could impact converting qualitative answers to quantitative metrics. To curate a knowledge graph of political data, annotators’ demographics information could influence their political stands, thus resulting in disputable annotations. Expertise level matters for domain-specific network annotations. For instance, domain experts need to perform biology experiments to annotate whether a protein-protein interaction network is an enzyme. The annotator demographics could help network dataset consumers assess the quality of the annotations and potential sources of errors.

5.3 Uses
This section allows network dataset consumers to quickly grasp what the dataset should and should not be used for. We include three parts: primary use and other uses. We understand that the use space of the network dataset could be huge, and it is impossible to list every use case. We encourage network dataset providers to list typical ones in this section.

Primary intended uses: This part details the use cases and tasks that the network dataset providers have in mind when creating the dataset. For machine learning datasets, this part should include the task type (e.g., classification, regression), task category (e.g., node, link or graph) and the evaluation metric(s) for the task. When there are no particular tasks for network dataset providers (e.g., A social network company releases part of the user network data to the public), the primary intended uses could be broad.

Other uses: This part details the potential uses cases of the network dataset. Examples include offline anomaly detection on a streaming network dataset, benchmarking HPC platforms on an online social network dataset, etc.

5.4 Network Statistics
This section includes point statistics and distributions of the network(s). The complex structure of networks makes it hard for people to understand whether the dataset is appropriate for their tasks by eyeballing the plain text description, a small subset (e.g., the egonet of one node), or one network plot. Network statistics could partially reveal such complexity with a few numerical values and visualizations of a few distributions. This section is inspired by the electronics industry, where standard characteristics (e.g., operating temperature) shown by numerical values and visualization of essential properties (e.g., time vs. frequency) are available in datasheets [46]. The electronics industry set such standards because small deviations that may seem insignificant can have severe consequences for the system as a whole [16]. However, the generative process of networks is even more complex, as the same configuration in graph models could result in significantly different networks. Similar to the datasheet in the electronics industry, we propose that a network report includes point statistics and visualizations. We list point statistics in Section 5.4.1, where each measure is a numerical value. We list distributions in Section 5.4.2, where visualization of distributions are shown.

Given the rich literature in network science, numerous statistical properties could be included in this section. We design the list through our formative interview, runtime concerns and diversity concerns. We prefer fundamental, well-studied measures over complex, domain-specific ones. Practically, network datasets could be massive (e.g., a billion nodes). Thus computationally heavy properties (yet important) like betweenness centrality and diameter are not included in the list, as both of the properties involve calculating the shortest paths of all pairs of nodes. We also prefer statistical properties that can cover diverse aspects of networks. Thus we choose point statistics and distributions that can cover (i) node and edge characteristics (size, degree, attributes); (ii) spectral properties (spectral radius, algebraic connectivity, singular value distribution, pagerank distribution); and (iii) network cohesion (connected components, average clustering coefficient, clustering coefficient distribution, degree assortativity coefficients, max k-core). Those point statistics and distributions could reveal fundamental phenomena of the underlying graph (e.g., the scale-free property, the small-world phenomenon).

Visualizations help network dataset consumers to understand distributions. Patterns in the visualization reveal the dynamics of the network. Abnormal patterns in the visualization (e.g., a spike) could help network dataset consumers understand the networks’ anomalies and further infer whether the anomalies would influence their tasks. We discuss common practices of visualizing network statistics in Section 5.4.2.

5.4.1 Point statistics. Point statistics map a network’s structure to a simple numerical space. For a multi-network dataset, network dataset providers can provide the average (and standard deviation) point statistics of all the networks in the dataset. Other point statistics beyond our list are also encouraged (e.g., structural balance for signed networks).

Number of nodes (|V|) and number of edges (|E|) describe the size of the network. Network dataset consumers can roughly
understand how much computation resources are necessary for processing the network.

**Proportion of nodes in the largest connected component** reveals the basic network connectivity. For directed networks, both weakly connected and strongly connected components should be considered.

Average degree ($\bar{d}$) is defined as $\bar{d} = \frac{1}{|V|} \sum_{u \in V} \deg(u)$, where the function $\deg(\cdot)$ denotes a degree measure (e.g., in degree) of a node. For undirected networks, $\bar{d} = \frac{2|E|}{|V|}$. For directed networks, the average out degree equals the average in degree, and both can be calculated as $d_{\text{out}} = d_{\text{in}} = \frac{|E|}{|V|}$. It quantifies the extent to which a node is connected to other vertices within the network.

(Tail) power-law exponent ($\gamma$) characterizes the degree distribution of the network. Many networks follow a degree distribution power law, i.e., the number of nodes with degree $n$ is proportional to the power $n^{-\gamma}$, for a constant $\gamma$ larger than one. There are multiple ways of estimating $\gamma$, we recommend estimating $\gamma$ using the robust method defined as $\gamma = 1 + \left| \left( \sum_{u \in V} \frac{\deg(u)}{d_{\text{min}}} \right)^{-1} \right|$, where $d_{\text{min}}$ denotes the minimal degree [42]. For networks that do not follow a power-law degree distribution, network dataset providers can indicate N/A for this item. N/A also informs network dataset consumers that the dataset is not scale-free.

Spectral radius ($\rho$) and Algebraic connectivity ($\alpha$) reveal mathematical properties on graph spectrum. Let $\lambda_1, \cdots, \lambda_n$ be the (real or complex) eigenvalues of $A$. Then $\rho = \max\{\lambda_1, \cdots, \lambda_n\}$. Algebraic connectivity $\alpha$ is the second smallest eigenvalue of $L$.

Average triangle count ($\bar{t}$) counts how many triangles one node is involved in. Triangles have been used in spam detection, link recommendation, and graph random graph models, etc [55]. We recommend to calculate it as $\bar{t} = \frac{1}{|E|} \sum_{(u,v,w) \in E} |\{ (u,v,w) \ | u, v, w \ \text{are connected} \} |$, because it is independent of the orientation of edges when the network is directed. Here

Average clustering coefficient ($\bar{c}$) quantifies the extent to which pairs of nodes with a common neighbor are also themselves neighbors. It extends the triangle count of a node by normalizing the triangle count with the number of pairs of nodes in its neighborhood. It is used as an indicator of "structural holes" in a network, which has rich applications in social networks [42]. As betweenness centrality (BC) is computationally intensive, average clustering coefficient can also be thought of as akin. Betweenness measures a node’s control over information flowing between all pairs of nodes in its component, while clustering coefficient is a local version of betweenness that measures control over flows between just a node’s immediate neighbors [42]. Mathematically, it can be calculated as $\bar{c} = \frac{1}{m} \sum_{u \in V} \frac{\deg(u)}{\deg(u)-1} - 1$.

Degree assortativity coefficient ($r$) quantifies the extent to which connected nodes have similar degrees. It is the Pearson correlation between the degree of connected nodes. Mathematically, it can be calculated as $r = \frac{\sum_{u,v \in E} (\deg(u)-\bar{d})(\deg(v)-\bar{d})}{\sum_{u \in V} \deg(u) \cdot \deg(v)}$. Besides, network dataset providers can also provide assortativity mixing by other measures of interests (e.g., Pearson correlation between the attribute of connected nodes). Some technologies related to networks such as graph convolution neural networks would fail if the network is not near homophily (i.e., connected nodes are similar) [27], thus providing one measure of homophily could help prevent misuse of those network technologies.

**Max k-core** ($k$) is the maximum $k$ such that $k$-core exists in the network, and a $k$-core is the maximal subgraph where all vertices have degree at least $k$. $k$-core has been applied to diverse areas such as hierarchical structure analysis, network visualization and network clustering [49]. Max $k$-core also provides the upper bound for max clique (i.e., another important structure in networks).

5.4.2 Distributions of node/edge statistics and attributes. We now discuss the distributions in network report and common design choices for visualization. For a multi-network dataset, network dataset providers can treat the dataset as a whole network, where each component is one network inside the dataset.

**Degree distribution ($\deg(\cdot)$):** The degree distribution is typically a skewed distribution, as there are more small-degree nodes than large-degree nodes. If the tail degree distribution follows a power law, then it’s typical to plot degree distributions on a log-log scale (i.e., both axes have a logarithmic scale) [42]. Histograms, line charts and point charts are common encoding choices. Histogram involves binning, and inappropriate binning can hide the trend in the data. For simplicity, we recommend point charts or line charts with degree on the x-axis and frequency (or probability mass function, or inverse cumulative distribution function $1$) on the y-axis.

**Pagerank distribution ($pr(\cdot)$) [45]:** Pagerank distribution can be thought of as an extended form of degree distribution, which considers how many neighbors a node has and how important those neighbors are. It can be calculated as $pr(G) = (1 - \alpha A D^{-1})^{-1}$, where $\alpha = 0.85$. Given its similarity to degree distribution, similar design choices in axes (e.g., log-log scale for scale-free networks) could be used. One significant difference is that pagerank is a continuous variable, while degree is a discrete variable, where the value is an integer at most the max degree. So we recommend point charts or line charts with pagerank value on the x-axis and inverse cumulative distribution function on the y-axis.

**Singular value ($\Sigma$) or eigenvalue ($\Lambda$) distribution:** Singular values (and eigenvalues) reveal the spectral properties of the graph. Spectral graph properties have broad applications in graph partitioning (clustering, community detection), dynamical systems [42]. There are variations of graph matrices ($A$, $L$, $\cdots$) and spectral decomposition (eigenvalue decomposition, singular value decomposition, $\cdots$) that characterize spectral properties. One common way to reveal the graph spectrum is to use the eigenvalues of the adjacency matrix [42]. However, for directed graphs, the eigenvalues are complex numbers, which brings additional difficulty in visualization. We recommend using the singular values of the adjacency matrix, which bypasses the complex number problem, and gives a unified view on both undirected and directed networks. Calculating the exact solution of all singular values is computationally expensive, so we recommend computing the approximate top-k singular values. The distribution can be plotted as a line chart or a point chart can be the encoding choices.

---

[1] Let $X$ be a random variable with cumulative distribution function (CDF) $F$. The inverse CDF is defined as $F^{-1}(q) = \inf\{x : F(x) > q\}$ [56].
Clustering coefficient distribution \((CC(\cdot))\): Clustering coefficient (i.e., local clustering coefficient) is the fraction of pairs of neighbors of node that are themselves neighbors. For all nodes \(u \in V\), \(CC(u) = \frac{\text{number of pairs of neighbors of } u \text{ that are connected}}{\text{number of pairs of neighbors of } u}\). Clustering coefficient distribution is closely related to the small-world phenomenon of real-world networks [57]. It is a continuous variable with values between 0 and 1. A common practice to visualize clustering coefficients as a cumulative distribution function on points or lines with linear scales [42].

Node attributes: For a quantitative attribute, a common practice is to use histograms. For an ordered attribute, a bar chart can be used to show the frequency of each category. For a categorical attribute, network dataset providers can use a bar chart to indicate the frequency of each category, and arrange the categories accordingly. Horizontal bar charts make it easier to have readable labels than vertical bar charts. If there are too many categories, one typical practice is to list top-k categories in the bar charts, and use ‘others’ to indicate the remaining categories.

Edge attributes: Visualization of an edge attribute is similar to that of a node attribute for a quantitative/ordered/categorical one. For temporal networks, where each edge contains a timestamp, we recommend a line plot with the relative time stamp on x-axis and number of edges within a fixed time window on the y-axis.

7 INITIAL USER STUDY

7.1 Study Design

We conducted an initial qualitative user study to explore the functionality of the network report in an academic environment. We hypothesized that network reports facilitate dataset communication and paper understanding.

Datasets and Paper We used the paper “Higher-Order Label Homogeneity and Spreading in Graphs”, and the POKEC dataset [12]. This version of POKEC dataset was built on top of the original POKEC dataset present in [52]. This mimics a common academic scenario, where datasets augment other datasets.

Study Protocol Each session began with a 15-minute tutorial on the main idea of the paper. We then asked participants to read the dataset description, followed by the network report in Figure 4. After that, we asked them to read through the experimental result section in the paper. During the study, we answered questions regarding the paper, but we did not answer questions regarding the dataset. Finally, we did an semi-structured checkout interview. Audio was recorded via Zoom.

Participants We recruited 6 participants (3 females, 3 males), all graduate students (1 MS, 5 PhD), with prior ML experience. 4/6 participants have research experience in networks. 2/6 participants have industrial experience related to network data. One of the participants had a hard time understanding the paper, so we discarded this session in analysis.

7.2 Analysis & Results

After transcribing the interviews, two of the authors did a line-by-line open coding on one transcript and produced a coding guide. Then the two authors coded separately on all transcripts, and discussed the codes to produce the results.
that documentation like network reports are important for building trust especially when the data is not publicly available.

7.2.3 Create a structured channel to communicate network datasets. All participants mentioned that the network report helps people communicate network datasets. P5 emphasized that it creates transparency, and said that "There are a thousand ways that I could construct a graph from [...] data. Whatever format I am deciding on, I should explain what I have done in detail and the rationale for why I have done it. Otherwise, it seems not genuine enough. And I think it is a great step towards more transparency in the network domain, which is needed. " 3/5 participants said that the network report allows for comparison and using data from other domains. 4/5 participants mentioned that the network report saves researchers’ time. P3 said that "Section 3 (Dataset Collection, Preprocessing and Annotation) is useful to me. ... I didn’t think of [network sampling and data filtering] until I saw it." P1 said that [the network report] saves time. ... And if you have a bunch of [network datasets], and you are deciding which dataset to use for your own research. And you can go to the uses section, and the data collection section, and find out which one is most specific to me. " 2/5 participants mentioned that they would open a Jupyter notebook to make some visualizations, so the section of Network Statistics is very useful. As participants find almost all sections of the network report useful, we conclude that the network report creates a structured way to communicate network datasets.

7.3 Limitations

The qualitative study was conducted in an academic environment, so the above results and analysis may bias towards academic use cases. As the user study is an initial study to explore the functionality and usability of the network report, it is not intended to be a complete evaluation of the network report.

8 DISCUSSION AND CONCLUSION

We have proposed the network report as a structured template to report information about network datasets. The network report includes network characteristics, information that captures population, bias and fairness, and detailed data preparation information. Our qualitative study suggests that the network report could help facilitate structured communication between network dataset providers and consumers.

Although network reports are designed to be flexible across different domains and use cases, the usefulness of network reports still relies on the integrity of the creators. In the near term, it is unlikely that the full contents of network reports could be automatically created and standardized to prevent all inappropriate use. We envision a standardized export formats and data loaders for networks accompanying network reports in the future. Besides, network reports are just one approach to increase transparency in network related technologies, which could include, for example, model cards for pre-trained graph neural network models [39] and vulnerability and robustness testing [15]. Future work includes accelerating the creation of network reports, studying how the information in network reports is interpreted and used by different network dataset consumers: researchers, developers, policy makers, etc, and how network reports could be integrated with other transparent tools.
[61] Jiaxuan You, B. Liu, Rex Ying, V. Pande, and J. Leskovec. 2018. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. In NeurIPS.