Summarization and Information Loss in Network Analysis*

Jamie F. Olson† Kathleen M. Carley†

Abstract

The increasing availability of large-scale network data makes the problem of network summarization especially relevant. In any data summarization, however, it is important to remain aware of the information not being presented. As such, we present a mathematical framework within which to consider the problem of network summarization. Using that framework, the concept of information entropy is applied in the context of network summarization. Finally, an O(—E—) algorithm is presented for computing information entropy in the case of simple deterministic network summarizations.

1 Introduction & Background

1.1 Motivation One of the many challenges network analysts face when confronted with real-world problems concerns the sheer size of real-world network data. Even a relatively small dataset is likely to contain thousands of nodes and many more connections. Particularly in exploratory analysis, analysts often turn to some variety of network summarization. This may consist of anything from analyzing a subset of the network to analyzing aggregate groupings of vertices as determined by some network grouping algorithm[1].

In general, the purpose of any network summarization is to focus on the information of interest to the research problem by setting aside some other information. Although network summarization is often a necessary step, it requires a balance between emphasizing some hopefully information and ignoring some hopefully unimportant information. There is always a risk that the information being ignored is actually the most important. By quantifying the precise amount of information being ignored, we hope to better enable network analysts to make informed decisions about the most appropriate level of summarization.

1.2 Problem Domain As an example of the ubiquitousness of network summarization, consider the problem of geospatially enable network analysis. As GPS devices becomes increasingly affordable and usable, the automated collection of geospatial information likewise becomes both technically and economically feasible. This is of particular importance in the application of dynamic network analysis to problems in law enforcement and counter-terrorism efforts. In these domains, spatially tagged network data has the potential to widen the scope of problems for which dynamic network analysis is feasible and to enable an expansion from strategic analysis into tactical analysis.

In order to achieve these goals, however, we need to develop an integrated approach to network and geospatial analysis. One simple step in the analysis of geospatially tagged data would be to simply overlay the network data onto a geospatial map. Figures 1(a) and 1(b) show two examples using this approach.

There is, however, a problem with this approach. By emphasizing the relationships between geospatial locations, we are discarding the the relationships within regions. We may see connections from A to B and B to C and infer a path from A to C through B, but there may in fact be no viable path through those regions. Of course, we can take steps to alleviate this particular dilemma, but the general problem of information loss still applies.

An appropriate measure of network information loss must have some way of being combined with geospatial information loss to provide an overall measure of the quality of the resolution.

Relational datasets of all types can be extremely large and network summarization is one simple method for analysts to quickly gain insight into the data. The proposed metric has a simple intuitive interpretation as the percentage of information lost at a particular resolution. By providing the amount of network information lost, we can help analysts choose the most appropriate resolution for analysis. In addition, by leveraging the general framework of information entropy, our proposed metric can be combined with measures of information

*This work was supported in part by the NSF Igert in CASOS (DGE-9972762), Air Force Office of Sponsored Research/ AFRL for the MURI with GMU on Cultural Modeling of the Adversary, 600322,ARO W911NF0710317 for work with ERDC-TEC and ONR N000140610921.Additional support was provided by CASOS and ISRI at Carnegie Mellon University. The views andconclusions contained in this document are those of the authors and should not be interpreted asrepresenting the official policies, either expressed or implied, of the Office of Naval Research, the ArmyResearch Office, the Department of Defense, the National Science Foundation, or the U.S. government.

†School of Computer Science, Carnegie Mellon University.

loss in other types of summarization(e.g. geospatial location).

2 The Model

We define a metric, *network entropy*, to be used as a way of measuring information gain/loss in networks. Information entropy is a way of quantifying the amount of information in terms of the certainty of the information. If we gain some information, E, the information gain(IG) of E is defined by (2.1-2.2).

$$IG[E] = h[X] - h[X|E] \tag{2.1}$$

$$h[X] = \sum_x p(x)\log(p(x)) \tag{2.2}$$

2.1 Definitions Let a graph consist of a set of vertices, $G = \{1, 2, ..., N\}$ and an affinity matrix, A, where the value of a_{ij} represents the connection between i and j. The size of the graph is represented by n.

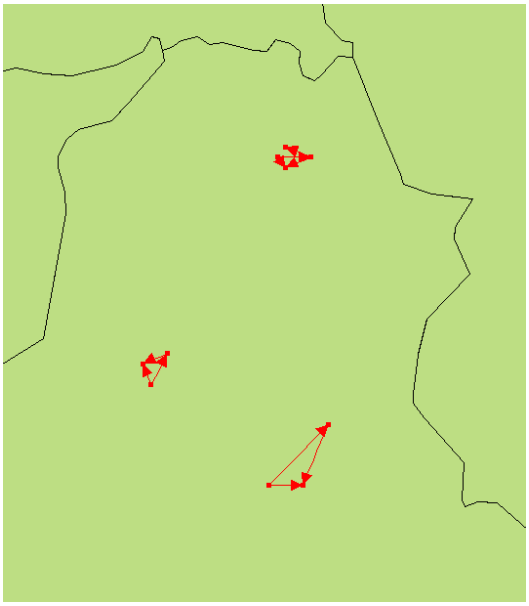
A graph summary can be any means of representing some subset of the information contained within that graph. Network partitioning and grouping algorithms are one popular example of graph summarization, but other methods are possible. The important characteristic of any network summarization method is that it implies, explicitly or implicitly, an approximation of the original network. In the case of a grouping algorithm, the original network is reduced to a network of groups, with the vertices within each group as irrelevant. We define this as a vertex reduction of a graph and the rest of the paper deals primarily with such summarizations. However, the network entropy can be calculated for any network summarization that produces an approximation of the network.

Let a node reduction of the graph, G, be defined by the $k \times n$ matrix, C, where $k < n$ is the reduced resolution. The reduction, $\acute{G} = \{1, 2, ..., k\}$, consists of k meta-vertices, where each meta-vertex, a, contains some set of vertices from the original graph. For each <meta-vertex,vertex> pair, $a \in C, i \in G$, define c_{ai} as $P(x \in a|C, x = i)$(hereafter $P(a|i)$). More intuitively, c_{ai} is the extent to which i is represented by a.

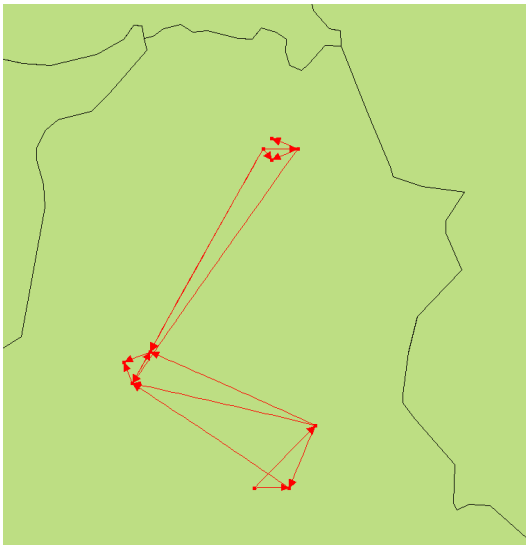
$$C = \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{k1} & \dots & c_{kn} \end{pmatrix}.$$

For convenience, define the variable p^C as a vector of probabilities, where $p_a^C = \sum_{i \in G} c_{ai}$. Now we can compute $P(i|a) = \frac{P(a|i)P(i)}{P(a)} = \frac{c_{ai} \times 1/n}{p_a^C}$.

In practice, many node reductions are simple deterministic reductions whereby each vertex in the original network is contained by a single meta-vertex with



(a) Network A



(b) Network B

Figure 1: Two simple networks with different geospatial properties

probability 1.0. In these node reductions, we define $c : G \rightarrow \acute{G}$ as the function that maps vertices in the original network to meta-vertices in the summary network.

2.2 Defining Entropy We claim that the information of interest in the network is the value of the edge between each pair of vertices. Therefore, we define the entropy of a network as the sum of the entropy of each edge in the network. In order to define the entropy of an edge, we must assume some underlying edge probability distribution. The most intuitive representation of an edge is as a Bernoulli random variable, but other probability distributions are certainly possible. Formally, we assume the edge for each pair, $< i, j >$, is an independent random variable X_{ij}. We assume that the original graph, G, is the ground truth, where $X_{ij} \sim Bernoulli(a_{ij})$. Each edge in the node reduced graph is a linear combination of the random variables in the original graph.

$$X_{ab} = \sum_{i,j \in G} P(i|a) \times P(j|b) \times X_{ij} \tag{2.3}$$

We can now create a maximum likelihood estimate of X_{ab} provided that we know the underlying distribution. If we assume $X \sim Bernoulli(p)$, we can calculate the parameter, p_{ab}.

$$\acute{a}_{ab} = \acute{p}_{ab} = \sum_{i,j \in G} P(i|a) \times P(j|b) \times a_{ij}$$

$$= \sum_{i,j \in G} \frac{P(a|i)P(i)}{P(a)} \times \frac{P(b|j)P(j)}{P(b)} \times a_{ij}$$

$$\acute{a}_{ab} = \sum_{i,j \in G} \frac{c_{ai}}{n \times p_a^C} \times \frac{c_{bj}}{n \times p_b^C} \times a_{ij} \tag{2.4}$$

We can then go backwards, reconstructing the approximation of the original graph that is implied in the node reduced graph.

$$\hat{a}_{i,j} = \hat{p}_{i,j} = \sum_{a,b \in \acute{G}} P(a|i) \times P(b|j) \times \acute{a}_{ab}$$

$$= \sum_{a,b \in \acute{G}} c_{ai} \times c_{bj} \times \acute{a}_{ab}$$

$$\hat{a}_{i,j} = \sum_{a,b \in \acute{G}} c_{ai} \times c_{bj} \times \sum_{i,j \in G} \frac{c_{ai}}{n \times p_a^C} \times \frac{c_{bj}}{n \times p_b^C} \times a_{ij} \tag{2.5}$$

We can then use these approximations to calculate the entropy of each pair of vertices, $i, j \in G$ as they are represented in the node reduced graph, $\acute{G}$. The additive property of entropy allows us calculate the entropy for the entire graph as the sum of the entropies for the edges.

$$H[i, j|C] = -\hat{p}_{ij} \log(\hat{p}_{ij}) - (1 - \hat{p}_{ij}) \log(1 - \hat{p}_{ij}) \tag{2.6}$$

$$H[G|C] = \sum_{i,j \in G} -\hat{p}_{ij} \log(\hat{p}_{ij}) \tag{2.7}$$

In general, we can compute the network entropy for any network summarization provided that we are willing to assume an underlying edge probability distribution and that we can reconstruct an approximation of the original network based on the summary network. For any network summary, $\acute{G}$.

$$H[G|\acute{G}] = -\sum_{i,j \in G} \int_x P(a_{ij} = x|\acute{G}) \tag{2.8}$$

2.3 Example Consider the simple network depicted in Figure 2(a) and the summary/partition of it depicted in Figure 2(b). The summary implies that the nodes A,B anc C are interchangeable and that D,E and F are similarly indistinguishable. The between-groups summary edge has a value of 1, with 9 possible pairwise combinations of vertices in the two groups. Therefore, the network summarization implies an approximation where each of the nine edges in the original graph (A→D,A→E,...C→E,C→F) has a value of 1/9. Assuming a Bernoulli distribution, we can calculate the entropy for each of those nine edges.

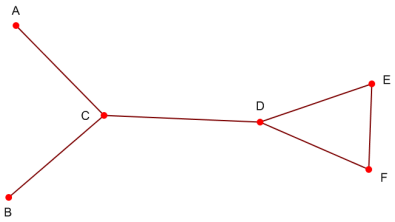
$$h[A \rightarrow B|\acute{G}] = -P(A \rightarrow B||\acute{G}) \times \log P(A \rightarrow B||\acute{G}) - (1 - P(A \rightarrow B||\acute{G}) \tag{2.9}$$

$$h[A \rightarrow B|\acute{G}] = 0.5032583 \tag{2.10}$$

We can similarly calculate the entropy of the edges within the vertices A,B,C and D,E,F as 0.9182958 and 0, respectively[1]. Using the additive property of entropy, we find that the entire network summary has an entropy of $9 \times 0.5032583 + 3 \times 0.9182958 + 3 \times 0 = 7.284212$. If we knew nothing about the network other than size and density, we would estimate each edge having a value of 0.4, with the entire network having an entropy of

[1] This assumes we ignore self-links. If we include self-links in the calculation and assume all vertices are self-linked, we find entropy values of 0.6500224 and 0.

14.56426. If knowing the entire network (by definition 0 entropy) results in reduction in entropy of 14.56426 and the network summary reduced entropy by 7.28 then we know that the network summary encapsulates 50.0% of the information in the network.



(a) Simple Network

(b) Partitioned Simple Network

Figure 2: Two simple networks with different geospatial properties

2.4 Calculating Entropy Algorithm 1 shows the naïve approach to computing entropy. As you can, this has a time complexity of $O(n^2k^2)$ and a space complexity of $O(n^2)$. However, most existing network grouping algorithms are not probabilistic and the computation becomes much simpler by taking advantage of this. Algorithm 2 takes as input a simple reduction function, $c : G \rightarrow \acute{G}$ and an edge set, E. Doing so reduces the space and time complexity to $O(|E|)$, where $|E|$ is the number of edges in the graph[2].

2.5 Importance of Entropy The use of information entropy as an underlying metric makes the integra-

[2]Assuming $|E|$ is much larger than the largest meta-vertex in the reduced graph.

input : G,E,C,p
output: The entropy, h of the graph G under the reduction, C

```
1  á ← 0
2  for i ← 1 to n do
3  |  for j ← 1 to n do
4  |  |  for a ← 1 to k do
5  |  |  |  for b ← 1 to k do
6  |  |  |  |  á [a][b] ← á [a][b] + C [a][i] ×
   |  |  |  |  C [b][j] × A [i][j]/(p [a] × p [b]
   |  |  |  |  ×n × n);
7  |  |  |  end
8  |  |  end
9  |  end
10 end
11 â ← 0  for i ← 1 to n do
12 |  for j ← 1 to n do
13 |  |  for a ← 1 to k do
14 |  |  |  for b ← 1 to k do
15 |  |  |  |  â [a][b] ← á [a][b] + C [a][i] ×
   |  |  |  |  C [b][j] × á [a][b];
16 |  |  |  end
17 |  |  end
18 |  end
19 end
```

Algorithm 1: A Vertex-based Algorithm

input : G,E,c,p
output: The entropy, h of the graph G under the simple reduction, c

```
1  á ← 0
2  for edge e in edges E do
3  |  i ←From(e)
4  |  j ←To(e)
5  |  a = c(i) b = c(j) á [a][b] ← á [a][b] + A
   |  [i][j]/(p [a] × p [b] ×n × n);
6  end
7  â ← 0  for edge é in edges É do
8  |  a ←From(é)
9  |  b ←To(é)
10 |  for node i ∈ G where c(i) = a do
11 |  |  for node j ∈ G where c(j) = b do
12 |  |  |  â [a][b] ← â [a][b] + á [a][b];
13 |  |  end
14 |  end
15 end
```

Algorithm 2: An Edge-based Algorithm

tion of network information gain with geospatial information gain fairly straightforward. We can apply information entropy to geospatial information gain as long as we are willing to assume some underlying probability distribution for the geospatial data as we did for the network data. A normal distribution may be the most intuitive, but other distributions may be more appropriate depending on the particular domain. We assume that the set of actual locations in each aggregated region is a sample of independent identically distributed random variables. We then use the sample of locations to estimate the distribution for the given region. For example, assuming a normal distribution, we estimate $\mu_x, \mu_y, \sigma x, \sigma_y$ as the sample means and standard deviations of the latitudes and longitudes. We can then compute the differential entropy of the region(2.11). In the case of a normal distribution, this reduces to (2.12)[2]. By using the additive property of entropy, we can compute the entropy of the entire dataset(2.13).

$$h[r] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y|r)\log(P(x,y|r))dxdy \tag{2.11}$$

$$h[r] = 1/2\log(2\pi\exp\sigma_x^2) \times 1/2\log(2\pi\exp\sigma_y^2) \tag{2.12}$$

$$h[R] = \sum_{r\in R} h[r] \tag{2.13}$$

Because entropy is additive, we can compute the overall information gain/loss simply by summing the network information entropy and the geospatial information entropy. This allows us to seamlessly integrate the two different dimensions of the data.

3 Discussion & Conclusions

We examine the idea of information gain as it applies to network summarization and propose information entropy as a way of measuring the quality of a resolution. We develop a probabilistic model of the information contained in a network and derive a formula for information entropy based on this model.

Although the computational complexity of computing network entropy depends on the network summarization method used, we develop an O(E) algorithm for simple deterministic node-reduction summarizations.

Network summarization has applications in a wide variety of network analysis problems. In these situations, network entropy can be used by analysts to determine the most appropriate level of summarization. By using information entropy as the basis for measuring the information in the network, we can combine the network information with information provided by other attributes (e.g. geospatial labels) in order to provide a comprehensive picture of the information contained in a particular network resolution.

References

[1] A. Clauset, M. E. J. Newman, and C. Moore, *Finding community structure in very large networks*, Phys. Rev. E, 70 (2004), p. 066111.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[3] S. Kullback and R. A. Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics, 22 (1951), pp. 79–86.

[4] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.