ABSTRACT

In a composite network, a piece of information traveling through links in a social network may have to travel over multiple links in an associated communication network. In this paper, we propose a model of composite networks that consists of two networks and an embedding between them, and several composite metrics that characterize information flow under a particular type of embedding. We present analytic results for the scaling behavior of “constrained composite stretch” of a path, “constrained composite diameter” of a graph, and “constrained composite broadcast time” of a tree, under random uniform embeddings onto various communication network structures. We validate our analytical results on composite stretch using two data sets consisting of a friendship social network geographically spread across Western Europe and a historical deployment of a military chain of command. We also present a randomized model of field deployment consistent with real-world data, and use simulations over this model to explore the distribution of constrained composite broadcast time. Finally, we show that our analytical bounds for composite broadcast time agree well with the simulation results.

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Categories and Subject Descriptors
G.2.2 [Discrete Mathematics]: Graph Theory—Network problems; Trees; Path and circuit problems; C.2.1 [Computer-Communication Networks]: Network Architecture and Design—Network topology; Wireless communication

General Terms
Theory, Networks

1. INTRODUCTION

The past few years have seen explosive growth of online social networks along with proliferation of the communication networks such as the Internet, wireline telephone networks, cellular telephone networks, and even multi-hop wireless ad hoc and sensor networks, in specialized application scenarios such as disaster relief and battlefield operations. Human beings usually share information with folks they know, and then that information gets forwarded along various links in the social network, either verbatim (e.g., directives from a leader flowing through the chain of command) or after modifications (e.g., propagation of rumors, gossip, or news from the field to the newspaper reader). This results in a social network constraining or guiding the spread of information according to its topology. Such constraints can result in a piece of information having to traverse much longer paths in the underlying communication network between the producer and the ultimate consumer of the information. This stretch is justified because either the intermediaries could play a critical role in interpreting/massaging the information or they serve as important links in the acquaintance chain, without whom the producer and the ultimate consumer would not know each other.

The total time taken for a piece of information to spread through the entire social network in the aforementioned constrained fashion is often different from the time taken to just multicast the information on the underlying communication network. Additionally, a piece of information may traverse a communication link several times during the process. While this is not an issue for lightweight content such as text, it can be a major issue for multimedia content, especially in an ad hoc network/DTN setting where accessing multimedia content directly from a server over a flaky network is not reliable.

In this paper, we present a systematic analytical study of the constrained information flow problem – in particular, we model a pair of networks (social and communication) as a composite graph – structures that result from embedding
the former network into the latter using mapping functions that can map a social network node to a communications network node when the former uses the latter as his/her communication portal.

Interactions between different genres of networks are becoming increasingly observable in both natural and man-made systems. Examples include peer-to-peer overlay networks [13], power grid networks and communication networks [5] or the interaction between the programmer social network and interconnected web of software modules developed by them [12].

We analytically study how the “constrained composite path stretch” and “constrained composite broadcast time” metrics scale with the sizes of both networks under consideration. One application scenario where this can be useful is a random deployment of workers or soldiers in a disaster relief site and their leader attempting to disseminate instructions and directives along his/her chain of command (essentially a social network of sorts). These messages trace a logical path in the social network which translates to a longer physical path on the underlying communication network.

1.1 Related Work

Related work in this area falls in three major categories: graph embedding, overlay networks, and network science approaches to study composite networks. Graph embedding has received attention in the parallel computing domain where the problem is to map a task graph onto a multiprocessor interconnection network (also known as host graph) [3, 14, 8], and in the ubiquitous computing domain where the problem is to map heterogeneous task graphs on non-regular networks such as mobile ad hoc networks [2], while attempting to determine the optimal mapping (or task to processor assignment) function such that metrics such as delay-to-task-completion, edge dilation (or stretch), node/edge congestion, etc. are minimized. Instead of the aforementioned “optimization” approaches, in this paper, we follow the “scaling law analysis” approach where both the graphs and the mapping function are given, and we characterize how a different set of appropriate “constrained” metrics such as composite path stretch, composite diameter, and broadcast time scale as a function of composite graph attributes.

There has been a significant amount of work on overlay networks in the past decade [13]. Seminal works such as CAN [15] and CHORD [16] attempt to design good distributed hash tables for P2P applications—the main endeavor there was to design good logical distributed data structures for storing (key,value) pairs overlaid on top of the Internet, so that efficient insertion and retrieval of hashed content is feasible from any part of the network. While this is a good example of a composite network, its similarities with our approach are slim. While overlay networks attempt to design good overlay graphs for the purpose of optimization of insertion/lookup overhead in our problem space, the social network graph is given, and we are interested in a different set of information flow metrics. Moreover, unlike the Internet which is a complete graph, our underlying network is a multi-hop network, in general.

Recently, network science researchers have begun studying various flavors of composite network structures. Kurant and Thiran propose the Layered Complex Network model [11] for studying load in transportation networks—they considered two-layer graphs where the physical graph models the transportation network and the logical graph models traffic flow between various cities; they use computational methods to determine levels of load on various transportation sectors in Europe. In comparison, our approach is analytical and we also study different metrics. A recent analytical line of research considers interdependent networks such as power grid and communication networks [5]—they use percolation theory to determine the fraction of nodes whose removal is likely to generate cascading failures in such networks. Leicht and D’Souza show that percolation thresholds of composite networks is lower than the individual networks, when considered separately [12]. While these approaches are all analytical, they study a different graph metric, i.e., degree of failure tolerance.

There is a large body of work pertaining to the embedding of one metric space into another—in particular, normed spaces such as d-dimensional Euclidean space $\mathbb{R}^d$—with “low-distortion”. This has been summarized well in [9]. This entails establishing the necessary and sufficient conditions on the properties of the two spaces for finding such embedding functions that yield a particular distortion, and in many cases finding the best embedding function [1]. A related idea of finding embeddings is popular in geographic routing—virtual coordinates are assigned to nodes in a hyperbolic space, and such an embedding guarantees that a greedy algorithm on the virtual coordinate space yields a route between every source and destination, if one exists [10].

The focus of this paper is not to find the best embedding function that yields a low distortion—rather, it is to analyze the distortion (or stretch) of an information flow that results from a random embedding of the nodes of the first graph onto the second graph, in distribution or in expectation.

1.2 Our Contribution

A summary of our contributions is as follows: 1) novel models and metrics for constrained information flow in composite networks; 2) mathematical analysis of scaling laws for constrained composite path stretch when a social network path is randomly mapped onto a general graph under both one-to-one and many-to-one mappings; 3) scaling laws for constrained composite broadcast time of a tree social network (chain of command) randomly mapped onto different communication networks; 4) validation of a subset of these results using the FOAF (friend of a friend) data set embedded on a geometric graph as well as a historical deployment of a chain-of-command.

2. COMPOSITE GRAPH MODELS

For two graphs $G_1$ and $G_2$, we define the composite graph $G$ to be the 3-tuple $(G_1, G_2, R)$, where $R \subseteq V(G_1) \times V(G_2)$ is an embedding relation between the vertex sets $V(G_1)$ and $V(G_2)$ of the two graphs, respectively.

In this paper, we focus on analyzing random embedding relations, where vertices in $G_1$ are mapped to vertices in $G_2$ via some random process $\pi$. In particular, we study two cases:

1. Each vertex in $G_1$ is mapped to a vertex in $G_2$ that has been sampled uniformly at random with replacement. This is the many-to-one scenario, where many “social network” nodes can get mapped to the same communication network node.

2. Each vertex in $G_1$ is mapped to a vertex in $G_2$ that has
been sampled uniformly at random without replacement. This is the one-to-one scenario, where a communication network node can host at most one social network node.

In general, every element of $R$ may have multiple attributes associated with it but in this preliminary study we only consider a binary relation. Note that this relation may be time-varying as an information object may be replicated or may even move from one communication node to another over time. This is a topic of future research.

We first analyze constrained composite path stretch, a metric that is useful for measuring how many physical communication hops are spanned by a logical information flow under a given embedding of the logical flow on a physical network.

Throughout this paper, let $G = (G_1, G_2, R)$ be a composite graph, with $V_i = V(G_i)$ the vertex set of graph $i$. Unless otherwise noted, $P_k = P_{uv} = \{u = v_0, v_1, ..., v_k = v\}$ is a path of length $k$ in $G_1$, and $d_{G_2} : V_2 \times V_2 \to \mathbb{R}$ is a shortest path distance metric in $G_2$.

**Definition 2.1** (cstretch). The constrained composite path stretch of $P_{uv}$ in $G$ is defined as:

$$\text{cstretch}_{G_2}(P_{uv}) = \sum_{i=0}^{k-1} \max_{x, y \in V_2} \{d_{G_2}(x, y)\}$$

CStretch characterizes the scenario with a stringent requirement that the information needs to traverse the nodes in the path $P_{uv}$ in order and in the process need to traverse the appropriately mapped nodes in $G_2$. This is not a farfetched scenario – in military systems, the chain-of-command (modeled by graph $G_1$) often mandates a piece of information to flow through the logical chain even though the ultimate recipient of the information may be in close proximity to the origin and the intermediate nodes are farther away from them. The reason behind this is that information often needs to get refined or obfuscated at each level of the logical chain before it is passed on further. Similarly, even in non-military applications (such as online social networks) information such as news or gossip is often routed along logical paths of friends who may be physically located all over the globe at large “Internet distances” from each other.

In the composite graph setting, the notion of diameter$^1$ can be extended to that of the constrained composite diameter which can be defined in terms of constrained composite path stretch.

**Definition 2.2** (ccd). The constrained composite diameter of $G$ is defined as

$$\text{ccd}(G) = \max_{u, v \in V_1} \text{cstretch}_{G_2}(P_{uv}).$$

The CStretch metric captures the extra distance in $G_2$ that a message has to travel in order to move through a path in $G_1$. We need a different metric to capture the combined stretch for a message traveling through a chain-of-command tree in a composite graph. In this context, it is more natural to consider the constrained composite broadcast time metric.

$^1$Diameter is the maximum length of the shortest path between any pair of nodes in a graph. It is an important measure for communication networks because it gives us a sense of the amount of time required (in the worst case) to traverse a network.

**Definition 2.3** (cbtime). Let $T$ be a tree in $G_1$, with root $u$. Then the constrained composite broadcast time of $T$ in the composite graph $G$ is defined as

$$\text{cbtime}_{G_2}(T) = \max_{v \in T} \text{cstretch}_{G_2}(P_{uv}).$$

The constrained composite broadcast time represents the stretch necessary to send a message through a chain-of-command tree that is deployed in a network topology. This may be of interest, for example, in a disaster relief situation when information needs to travel from a central director to end caregivers while relief workers are deployed in the field. In other words, it measures the time at which the last worker received the message that was broadcast through the chain of command.

3. COMPOSITE STRETCH ANALYSIS

In this section, we characterize the distribution of the constrained composite path stretch of $P_i$ over uniform random embeddings into $G_2$. We first prove some general results that apply to any graph $G_2$, and then illustrate scaling laws for a few well-known graph families.

3.1 Theoretical Results

For any graph $G = (V, E)$, let $D_G$ be the geodesic graph distance matrix between all pairs of vertices $v_i, v_j \in V$. That is, each entry $d_{ij}$ in $D_G$ represents the shortest path distance from $v_i$ to $v_j$ in $G$. Then we note that the sum of the geodesic distances $\Delta_G = \sum_{v_i, v_j \in V} d_{ij}$, is a constant depending only on the structure of $G$.

**Lemma 3.1.** Let $G$ be a graph with $|V| = n$, and let $X$ be a random variable denoting the geodesic distance between two vertices of $G$ chosen uniformly at random. Then:

$$E[X] = \begin{cases} \frac{\Delta_G}{n(n-1)}, & \text{when sampling without replacement} \\ \frac{\Delta_G}{n^2}, & \text{when sampling with replacement} \end{cases}$$

**Proof.** The case where sampling is done with replacement is clear: since there are $n^2$ pairs of vertices from which to choose, the expression given is the average distance. If sampling is done without replacement, then $\Delta_G$ double counts the distance for each of the $\binom{n}{2}$ unique pairs of vertices. Note that the $n$ diagonal entries in $D_G$ contribute nothing to $\Delta_G$.

**Corollary 3.1.** There is no asymptotic difference in $E[X]$ between sampling vertices with or without replacement.

**Proof.** From the preceding Lemma, it follows that the ratio of $E[X]$ when sampling without replacement to $E[X]$ when sampling with replacement is $1 + \frac{1}{n} \to 1$ as $n \to \infty$.

Next, we show that the expected stretch of a link is independent of the choices of vertices already mapped, regardless of whether sampling is done with or without replacement.

**Lemma 3.2.** Let $v_1, v_2, ..., v_i$ be a sequence of vertices chosen uniformly at random from $V$ (with or without replacement), and let $X_i$ be the random variable giving the distance between $v_i$ and $v_{i-1}$. Then $E[E[X_{i+1}|v_1, v_2, ..., v_i]] = E[X_2]$.

**Proof.** While the statement may be obvious for the case of sampling with replacement, we exercise more care for the case where sampling is done without replacement, and prove...
the statement combinatorially. For the RHS, select one vertex uniformly at random and color it red (call it \( v_1 \)). Then select another from the remaining and color it blue (\( v_2 \)). The RHS counts the expected distance between these two vertices. We now argue that the LHS counts the same. To see this, first color one vertex blue (call it \( v_{i+1} \)), and another vertex red (\( v_i \)). Now color \( i-1 \) other vertices green (\( v_{i-1}, \ldots, v_1 \)). The LHS counts the expected distance between the blue vertex and the red vertex.

This leads us to a general theorem about the expected composite stretch of a path.

**Theorem 3.1.** For a path \( P_k \) embedded uniformly at random into any graph \( G_2 \) (with the sampling performed with or without replacement),

\[
E[\text{stretch}^n_{G_2}(P_k)] = k \cdot E[X],
\]

where \( X \) is the random variable giving the distance between two randomly chosen vertices in \( G_2 \).

We emphasize the expectation is being taken over the uniform random embedding \( R_e \). But as we saw in Lemma 3.1, for a specific \( G_2 \), if the sampling method of \( R_e \) is known, then the expected distance \( E[X] \) is a constant.

**Composite Diameter:** In addition to the average-case, we also want to describe the worst-case \( \text{stretch} \) for a random embedding. It is easy to see that if \( R_e \) samples vertices with replacement, then each successive link in any path can simply bounce back and forth between the furthest two vertices in \( G_2 \). Thus, \( ccd(G) = \text{diam}(G_1) \cdot \text{diam}(G_2) \). However, when \( R_e \) samples vertices without replacement, the problem is an instance of MAX-TSP, which is MAX SNP-hard [7]. However, a greedy approximation heuristic works well in practice.

### 3.2 Examples

Theorem 3.1 shows that the expected stretch of a path is equal to the length of the path times a constant depending only on the structure of \( G_2 \) and the distribution of the random embedding. In what follows, we present examples of some well-known graph families, and illustrate how their structure affects the distribution of \( \text{stretch} \).

**d-dimensional Discrete Lattice:** Let \( D^d_n = \{0,1,\ldots,n-1\}^d \) be the \( d \)-dimensional discrete lattice on \( n^d \) points, and consider a composite graph with \( G_2 = D^d_n \). On this graph topology, geodesic distance is equivalent to the \( l_1 \)-norm (Manhattan distance) between two points in \( D^d_n \). Thus, \( dG_2(v,w) = \sum_{i=1}^{d} |v_i - w_i| \), and summing all \( n^{2d} \) of these pairs gives

\[
\Delta_{G_2} = \sum_{v,w \in V} \sum_{i=1}^{d} |v_i - w_i| = \frac{dn^{2d+1}}{3} \left(1 - \frac{1}{n^2}\right)
\]

It follows from Lemma 3.1 and Theorem 3.1 that under a random uniform embedding with replacement into the \( d \)-dimensional discrete lattice,

\[
E[\text{stretch}^n_{G_2}(P_k)] = \frac{kdn}{3} \left(1 - \frac{1}{n^2}\right)
\]

Note that in this case it is also straightforward to fully explicate the distribution of \( X \). For any \( 1 \leq i \leq d \), let \( X_i = |v_i - w_i| \). Then the probability mass function for \( X_i \) is

\[
p_{X_i}(\delta) = \begin{cases} \frac{1}{2(n-\delta)} & \text{if } \delta = 0 \\ \frac{1}{n^2} & \text{otherwise} \end{cases}
\]

since each coordinate can take on any of \( n \) values, and there are \( n-\delta \) ways to achieve each value of \( \delta \) between 0 and \( n-1 \). Since the \( X_i \)'s are independent and identically distributed, we can extract (among other things), the second moment of \( X \):

\[
\text{Var}[X] = d \cdot \frac{(n^2-1)(n^2+2)}{18n^2}.
\]

We can infer from this that the expected stretch is not likely to deviate from its mean.

For the discrete lattice, we have that \( \text{diam}(G_2) = d(n-1) \), so as mentioned above, the ccd for \( P_k \) is \( k(n-1) \). For the non-trivial “without replacement” scenario, we implemented a Greedy approximation heuristic, and verified that ccd for both without and with replacement scenarios are \( O(n^2) \).

**Cycle:** Let \( C_n \) be the cycle of length \( n \), and consider uniform discrete mappings from \( P_k \) onto \( C_n \). Clearly, the maximum distance between two vertices in \( C_n \) is \( \lceil \frac{n}{2} \rceil \). But, for each possible distance \( x \) between 0 and \( \frac{n}{2} \), there are exactly \( n \) such pairs for \( x = 0, 2, \ldots, \frac{2n}{2} \), and exactly \( 2n \) such pairs otherwise. It is thus straightforward to show that

\[
\Delta_{C_n} = \begin{cases} \frac{n(n+1)}{2} & \text{if } n \text{ is odd} \\ \frac{n^2}{4} & \text{if } n \text{ is even} \end{cases}
\]

Application of Lemma 3.1 and Theorem 3.1 then reveal that for random uniform embeddings onto \( C_n \),

\[
E[\text{stretch}^n_{C_n}(P_k)] = k \cdot \left( \frac{n^2}{4} + o(1) \right).
\]

Greedy is optimal on \( C_n \), since if \( n \) is odd, it finds \( n-1 \) pairs at distance \( \lceil \frac{n}{2} \rceil = \text{diam}(G_2) \) from each other, which is optimal by definition. On the other hand, if \( n \) is even, it picks all \( \frac{n}{2} \) pairs at distance \( \frac{n}{2} = \text{diam}(G_2) \) from each other, and another \( \left( \frac{n}{2} - 1 \right) \) pairs at the next greatest distance \( \left( \frac{n}{2} - 1 \right) \).

**Balloon graph:** Next, we consider a graph family with some interesting properties. Let \( B_{n,m} \) be a balloon graph consisting of a string (line graph) of length \( m \), connected to a balloon (clique) of size \( n - m \), for any \( 0 \leq m < n \). For clarity, we specify that vertices \( \{v_1, \ldots, v_m\} \) make up the string, while vertices \( \{v_{m+1}, \ldots, v_{n-1}\} \) make up the balloon. Note that for any two indices \( 0 \leq i < j \leq n-1 \) in this graph, we have that

\[
d_{B_{n,m}}(v_i, v_j) = \begin{cases} j - i & \text{if } i < j < m \\ m + 1 - i & \text{if } i < m \leq j \\ 1 & \text{if } m \leq i < j \end{cases}
\]

In particular, note that \( \text{diam}(B_{n,m}) = m + 1 \). In computing the distance matrix, we distinguish three cases based on the indices of the two vertices chosen:

1. If \( i \leq j \leq m \), then both vertices lie in the string, which is \( D^2_{m+1} \). This contributes \( \Delta_{D^2_{m+1}} \) towards \( \Delta_{B_{n,m}} \).

2. If \( m \leq i \leq j \), then both vertices lie in the balloon, and it is clear that on the complete graph \( K_n \), \( \Delta_{K_n} = n^2 - n \), since every pair of vertices are connected by an edge, but there are \( n \) ways to choose the same vertex twice.
3. If \( i < m < j \), then one vertex lies in the string, and the other lies in the balloon. Consider any vertex \( w_j \) in the balloon. Its distance from the set of vertices in the string is simply \( m + 1, m, m - 1, \ldots, 2 \). Thus, the contribution to \( \Delta B_{n,m} \) is
\[
2(n - m - 1) \sum_{i=2}^{m+1} i = m(m + 3)(n - m - 1).
\]

Adding these three quantities yields
\[
\Delta B_{n,m} = -\frac{2}{3} m^3 + (n - 2)m^2 + \left( n - \frac{4}{3} \right) m + n^2 - n.
\]

The reader may verify that setting \( m = 0 \) corresponds to the special case where the balloon graph is itself a clique, while setting \( m = n - 1 \) yields the special case where \( B_{n,n-1} = D_n^0 \).

By Theorem 3.1 and Lemma 3.1, the expected stretch for a path of length \( k \) onto \( B_{n,m} \) is thus:
\[
E[\text{stretch}_RGG(B_{n,m})] = \kappa(n) \cdot \left(1 + O\left(\frac{m^2}{n}\right)\right).
\]

Random Geometric Graph: Lastly, we consider the composite stretch when \( P_k \) is mapped onto a random geometric graph \( G_2 = RGG(n,r(n)) \), where \( r(n) \) is the radius of communication. That is, \( G_2 \) consists of \( n \) vertices placed uniformly at random in \([0,1]^2\), wherein any two vertices are connected with an edge if and only if the Euclidean distance between them is at most \( r(n) \). Gupta and Kumar [6] showed that a radius of connectivity of \( r(n) = \sqrt{\frac{\ln n + c(n)}{n}} \) ensures asymptotic connectivity in the RGG with high probability if and only if \( c(n) \to +\infty \). In all of our discussions on RGG in this paper, we assume that the radius of connectivity is at least this large, i.e., \( r(n) = O\left(\sqrt{\ln n/n}\right) \).

As before, Theorem 3.1 still applies, so it remains only to characterize the distribution of the random variables \( X \) giving the geodesic distance between two vertices in \( RGG(n,r(n)) \) selected uniformly at random. Note that in contrast to the previous examples we have considered, we now have two sources of randomness: 1) the randomized construction of the RGG; and 2) the random uniform embedding. If the Euclidean distance between two vertices in an RGG is \( \delta \), then recent results confirm that with high probability, the geodesic distance \( X \) differs from its minimum of \( \delta/r \) by at most a constant [4].

**Theorem 3.2.** With high probability, the expected geodesic distance in \( RGG(n,r(n)) \) satisfies
\[
\frac{\Delta(2)}{r(n)} \leq E[X] = \kappa(n) \cdot \frac{\Delta(2)}{r(n)},
\]
where \( \Delta(2) \approx 0.5214054331 \) is a known constant, and \( \kappa(n) \geq 1 \) is \( O(1) \).

**Proof.** Let \( u, w \) be two vertices in \( RGG(n,r(n)) \) selected uniformly at random, and set \( \delta = ||u - w||_2 \). Clearly, \( X \geq \delta/r \). Conversely, if \( \delta = \Omega(\log^{3.5} n/r^2) \), then by a result from [4], \( X = O(\delta/r) \).

Taking expectation yields the result, since \( E[\delta] = \Delta(2) \) is a known result [17]. \( \square \)

One metric of interest is the expected composite stretch when \( \phi(B_{n,m}) \) is the maximum (out)degree, \( \delta \), for some \( 1 \leq \delta < k \). We assume that \( T_k \) exists in some \( G_1 \), and examine the constrained composite broadcast time for sending a message from the root to each of the other nodes.

**Star Topology:** We begin with the special case where \( T_k \) is a \( k \)-star. First, we introduce some notation. Let
\[
p_k = \frac{1}{\binom{n-1}{k}} \left(0, \ldots, 0, 1, \begin{pmatrix} k \end{pmatrix}, \ldots, \begin{pmatrix} n-2 \end{pmatrix}, \begin{pmatrix} k-1 \end{pmatrix}, \begin{pmatrix} k \end{pmatrix} \right) \in \mathbb{R}^n
\]

**Table 1: Summary of Path Stretch Metrics for Uniform Random Embeddings of \( P_k \)**

<table>
<thead>
<tr>
<th>( P_k )</th>
<th>( \Omega(2) )</th>
<th>( E[\text{stretch}] )</th>
<th>( \max[\text{stretch}] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_n^0 )</td>
<td>( (1 - n^{-2}) )</td>
<td>( kd(n-1) )</td>
<td></td>
</tr>
<tr>
<td>( C_n )</td>
<td>( k \cdot \left[ \frac{\delta}{r} + o(1) \right] )</td>
<td>( k \cdot \left[ \frac{\delta}{r} \right] )</td>
<td></td>
</tr>
<tr>
<td>( B_{n,m} )</td>
<td>( k \cdot \left(1 + \Omega\left(\frac{m^2}{n} \right)\right) )</td>
<td>( k(m+1) )</td>
<td></td>
</tr>
<tr>
<td>( RGG(n,r(n)) )</td>
<td>( O\left(k \sqrt{\frac{n}{\ln n}} \right) )</td>
<td>( \in \mathbb{R}^n )</td>
<td></td>
</tr>
</tbody>
</table>

**Corollary 3.2.** For \( r(n) \) sufficiently large (i.e., greater than the critical connectivity threshold), the composite stretch of a path \( P_k \) on a random geometric graph \( RGG(n,r(n)) \) satisfies with high probability:
\[
E[\text{stretch}_RGG(P_k)] = k \cdot \kappa(n) \cdot \frac{\Delta(2)}{r(n)} = O\left(k \cdot \sqrt{\frac{n}{\ln n}} \right).
\]

**3.3 Average vs. Worst-Case Analysis**

Thus far, we have characterized both the average case (expected \( \text{stretch} \)) and the worst case (\(ccd\)) for a random uniform embedding of a path onto several graph families. For both the lattice and the cycle, these quantities were of the same order of magnitude. At this point a natural question arises: Are there graphs for which the ratio of the maximum \( \text{stretch} \) to the average \( \text{stretch} \) of \( P_k \) is not \( O(1) \)? Indeed, the balloon graph is one such graph. As the diameter of \( B_{n,m} \) is \( m+1 \), the maximum stretch is \( \text{diam}(G_1) \cdot (m+1) \). If we let \( \phi(B_{n,m}) \) be the ratio of the maximum \( \text{stretch} \) to the mean \( \text{stretch} \), we can see that:
\[
\phi(B_{n,m}) = \frac{\text{diam}(G_1)(m+1)}{\text{diam}(G_1)\left(1 + O\left(\frac{m^2}{n} \right)\right)} = O\left(\frac{n}{m}\right)
\]

In particular, then, for \( m = \sqrt{n} \), the ratio of the maximum stretch to the mean stretch for the balloon graph \( B_{n,m} \) is \( O(\sqrt{n}) \). Explicit calculations reveal that for \( m = \sqrt{n} \), in fact \( E[X] \to 2 \) as \( n \to \infty \).

More interesting is the fact that this gap appears to be mainly an artifact of the difference between sampling with and without replacement. The results of our Greedy algorithm for CCD without replacement suggest that with \( m = \sqrt{n} \), the CCD and expected \( \text{stretch} \) are of the same order to magnitude.

Table 1 summarizes our theoretical results.

**4. COMPOSITE BROADCAST TIME**

In this section, we analytically characterize the expected composite broadcast time for tree topologies. Social networks for information dissemination commonly have tree structures (more on this in section 5), hence this analysis can be useful for specific communication network deployment scenarios. Let \( T_k \) be a \( k \)-node tree of height \( h \) and maximum (out)degree \( \delta \), for some \( 1 \leq \delta < k \). We assume that \( T_k \) exists in some \( G_1 \), and examine the constrained composite broadcast time for sending a message from the root to each of the other nodes.
be a column vector, and note that \(|p_k|_1 = 1\). The \(i^{th}\) entry in \(p_k\) represents the probability that the \(i^{th}\) largest among \(n\) values is returned, when this value is the maximum among a subset of size \(k\) chosen uniformly at random. Furthermore, let \(f : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}\) be the function that sorts the rows of a matrix in ascending order from left to right. That is,

\[
D = \begin{bmatrix}
d_1 \\
d_2 \\
\vdots \\
d_m
\end{bmatrix} \Rightarrow f(D) = \begin{bmatrix}
\text{sort}(d_1) \\
\text{sort}(d_2) \\
\vdots \\
\text{sort}(d_m)
\end{bmatrix},
\]

where \(d_i\) is the \(i^{th}\) row of \(D\). Finally, \(v_m = \frac{1}{m}(1, \ldots, 1) \in \mathbb{R}^m\).

**Theorem 4.1.** For any graph \(G_2\), the broadcast time of a star of size \(k\) satisfies

\[
E[c\text{btime}_{G_2}(S_k)] = v_m^T \cdot f(D_{G_2}) \cdot p_k.
\]

**Proof.** Let \(d_i\) be the \(i^{th}\) row of \(D_{G_2}\), and suppose that the root of the star \(S_k\) is mapped to node \(i\) in \(G_2\). The broadcast time of \(S_k\) is the maximum \(c\text{stretch}\) from among its \(k\) children. But since the \(j^{th}\) entry of \(p_k\) is the probability that the \(j^{th}\) largest value in \(d_i\) will be returned, the inner product \(\langle \text{sort}(d_i), p_k \rangle\) gives the expected value of the maximum of the \(k\) \(c\text{stretch}\)es. Multiplication on the left by \(v_m^T\) simply averages these values over all \(n\) rows. \(\square\)

Note that this is consistent with Theorem 3.1 for the special case where \(k = 2\). Theorem 4.1 allows us to compute the broadcast time of a \(k\)-star for a variety of graph families, and we later use these as building blocks for bounds on general trees. Moreover, Theorem 4.1 improves on the trivial upper bound of \(\text{diam}(G_2)/2\) for the broadcast time of a star. A better bound can be derived by considering the average eccentricity of \(G_2\). The eccentricity \(\varepsilon\) of a vertex in a graph is defined as the maximum geodesic distance between that vertex and any other.

**Corollary 4.1.** For any graph \(G_2\), the broadcast time of a star of size \(k\) satisfies

\[
E[c\text{btime}_{G_2}(S_k)] \leq \frac{1}{n} \sum_{v \in V_2} \varepsilon(v)
\]

**Proof.** Substituting \(p_{n-1}\) in place of \(p_{k-1}\) returns the average eccentricity of the vertices in \(G_2\). \(\square\)

Corollary 4.1 provides a better bound than the diameter, but is not nearly as good as when using Theorem 4.1 directly. To illustrate how Theorem 4.1 can be used for a specific \(G_2\), we provide an upper bound on the broadcast time of a star, when \(G_2\) is the line lattice above.

**Corollary 4.2.** For \(G_2 = D_n^k\), the line lattice, the broadcast time of a star of size \(k\) satisfies

\[
E[c\text{btime}_{G_2}(S_k)] \leq \frac{k}{k+1} \cdot n.
\]

**Proof.** The maximum product on the right certainly oc-

\[
\langle d_1, p_k \rangle = \frac{1}{\binom{n-1}{k}} \sum_{j=k+1}^n \binom{j-2}{k-1}
\]

\[
= k \binom{n}{k+1} \frac{n}{k+1} = \frac{k}{k+1} \cdot n.
\]

\(\square\)

**Tree topology:** For any tree \(T_k\) with maximum degree \(\delta\), let \(\delta_i\) be the maximum outdegree among nodes at height \(i\). The eccentricity \(\varepsilon\) of a path of length \(h\), which is the longest in \(T_k\). No other single path from the root to a leaf could have expectation longer than this, so the expectation for the tree must be at least this large. \(\square\)

**Observation 4.1 (Ctime: Lower Bound).**

\[
E[c\text{btime}_{G_2}(T_h)] \geq \frac{h}{\max_{1 \leq \ell \leq h} \delta_i}
\]

**Proof.** The lower bound represents the expected \(c\text{stretch}\) of a path of length \(h\), which is the longest in \(T_k\). No other single path from the root to a leaf could have expectation longer than this, so the expectation for the tree must be at least this large. \(\square\)

**Observation 4.2 (Ctime: Upper Bound).**

\[
E[c\text{btime}_{G_2}(T_h)] \leq \sum_{\ell=1}^h E[c\text{btime}_{G_2}(S_{\delta_i})]
\]

**Proof.** The upper bound represents the sum (over all \(h\) levels of \(T_k\)) of the expected composite broadcast time for the largest star graph \(S_{\delta_i}\) at each level \(\ell\). This the maximum expected broadcast time, since no path from the root to a leaf could take longer than this. \(\square\)

Combining Theorem 3.1 with Observations 4.1 and 4.2 yields the following bounds on the expected broadcast time of a general tree.

**Corollary 4.3.** For any tree \(T_k\) of height \(h\) and maximum outdegree \(\delta\),

\[
h \cdot E[X] \leq E[c\text{btime}_{G_2}(T_h)] \leq h \cdot E[c\text{btime}_{G_2}(S_{\delta_i})],
\]

where \(X\) is the r.v. giving the expected geodesic distance between two vertices in \(G_2\).

**Proof.** The lower bound is an application of Theorem 3.1 to Observation 4.1, while the upper bound follows from Observation 4.2 and the fact that \(\delta = \max_{1 \leq \ell \leq h} \delta_i\). \(\square\)

## 5. Simulation Based Evaluation

In this section, we use a variety of simulation-based approaches to study how the composite stretch and composite broadcast time metrics behave under various choices of real (or at least realistic) social and communication networks. We first study the case where the social network (or \(G_1\)) is a hierarchical network like a chain of command that exists in military missions or in disaster relief operations. Our second case study is that of a social network that is richer than a tree as in friendship relationships.
5.1 Chain-of-Command

Figure 1(a) denotes the chain of command hierarchy from within a representative brigade reporting structure in the US Army. Commands move from the highest ranked (root, shown in pink) node to the lower ranked nodes in the tree until they reach the lowest ranked (leaf, shown in blue) nodes. Though not depicted in Figure 1(a), each node in $G_1$ occupies a physical location in space. In Figure 1(b), we show the geometric graph $G_2$ constructed from these physical locations by adding a communication edge between two nodes if they lie within a prescribed radio transmission range (i.e., the critical radius of connectivity $r(n)$ outlined above) of each other. The coordinates of the nodes in $G_1$ are specified according to a historical deployment scenario over a 124 $\times$ 148km$^2$ area. $G_2$ is not an RGG but only a geometric graph with radius of communication 14.93km.

The composite graph $G^* = (G_1, G_2, R)$ that combines Figures 1(a) and 1(b), along with the identity mapping $R$, is a realistic composite network structure for a military or disaster relief deployment. In this instance, the broadcast time is 18, though eccentricity in $G_2$ of the root node of $G_1$ is 9. Thus, the constraints imposed on the information flow by the chain of command require a message to travel through twice as many hops as was mandated by the actual deployment. A path that produces the broadcast time is highlighted in red in Figures 1(a) and 1(b).

In order to put this broadcast time of 18 in context, we simulated three different randomized scenarios, each of which could produce $G^*$ as a singular outcome:

- **S1** Instead of $R$ being the identity mapping, $R$ is a random permutation.

- **S2** Instead of $G_2$ being a geometric graph over the actual coordinates of deployment, each node in $G_1$ was assigned a random 2D coordinate drawn from a square region of the same area as in S1. Therefore, $G_2$ is a random geometric graph (RGG). Fig. 1(c) shows an example of such a deployment.

- **S3** Instead of $G_2$ being a geometric graph over the actual coordinates of deployment, the coordinates were generated according to a random model that is a function of the chain of command tree $G_1$, with $R$ as the identity mapping. Details of the model are given below.

In the actual deployment, note that the lowest ranked nodes are collocated in bunches of four, hence Fig. 1(b) appears sparser than Fig. 1(c). Moreover, the broadcast time jumps from 18 to 51 for this particular RGG. This is because in the real deployment, there is strong correlation between the location of a node and its rank in the command hierarchy (even though the maximum stretch is as high as 18), which does not exist in random deployments. Inspired by this observation, we constructed a correlated random deployment model for S3.

**Details of Deployment Model for S3:** Our model places each child in an equi-spaced, but randomly-oriented, ring around its parent, with a random jigger applied in both the horizontal and vertical directions. This process is recursively applied down the chain of command tree, which we assume has height $h$. Let $\nu_i$ be the node in $G_1$ at distance $h_i$ from the root node, and having $n_i$ children. Then the location of $\nu_i$'s children are determined as follows:

1. Find the $n_i$ roots on unity $\omega_1, \ldots, \omega_{n_i}$ and associate one with each of the $n_i$ children.

2. Draw a uniform random variable $u \in [0, 1]$.

3. Set the target distance from parent to child to be $\rho_i = a(h - h_i)^2$, where $a$ is a parameter determined from analysis of the actual deployment data. [In our case $a = 1.85$.]

4. For each $j \in 1, \ldots, n_i$, set the target location $c_j = \rho_i \cdot e^{2\pi i u} \cdot \omega_j$, and draw two random coordinates $x_j$ and $y_j$ from normal distributions with mean $\Re(c_j)$ and $\Im(c_j)$, respectively, and standard deviation $b \bar{h}_{\nu_i}$. Here $b$ is a parameter determined from the data (we assume a model with a constant coefficient of variation, $b = 0.293$), and $\bar{h}_{\nu_i}$ is the mean distance between parent and child at height $h_i$.

5. Return $(x_j, y_j)$.
Fig. 2(a) plots the simulated distribution of broadcast time for each of the aforementioned scenarios. That the red vertical line (corresponding to $G^*$) falls within the distribution of S3 (more precisely, in the 92nd percentile), suggests that our correlated deployment model is a useful one. Moreover, the other two scenarios, both of which were agnostic to the structure of the chain of command tree, performed comparatively poorly. Empirically, the probability of obtaining a broadcast time as low as that of $G^*$ via either S1 or S2 appears to be negligible. This illustrates the potential efficiency dangers inherent to composite networks.

5.2 Friend-Of-A-Friend (FOAF)

Now we consider a small social network data set that was extracted from the Semantic Web Billion Triples Challenge (BTC) program. The data contains unique identifiers and indicates the existence of friendship relations between pairs of users – as per the friend of a friend (FOAF) ontology. It also contains the geographic coordinates of users.

One can imagine the IDs in the data set communicating with their friends over some underlying communication network. In this scenario, the communication would typically happen over the wired Internet that connects various users on the map; however, we used this node distribution data to study the composite stretch of the FOAF social network on a geographically distributed multi-hop network assuming a geometric graph model as described in Sec. 5.1 — in particular, we place a node at each location in the data set and construct a graph using a transmission radius that is large enough to connect most nodes in the network (akin to the notion of critical radius in case of a random geometric graph). In Figure 3, we show a 237 node social network ($G_1$), alongside the geometric graph imposed over the same set of vertices using the connectivity radius $r(n) = 5.1$ degrees of latitude/longitude.

Since $G_1$ is a rich social network, all nodes can act as sources of information which can flow along random spanning trees of $G_1$. Hence we pick all nodes in $G_1$ one by one; and assign them as roots of random spanning trees. We then measure $cbtime$ for each root node and plot the results in Fig. 2(b) (black jitter-spaced hollow circles). The dashed black line shows the average $cbtime$ as a function of the height of the spanning tree. That this is approximately linear accords with Corollary 4.3.

On the same figure (Fig. 2(b)) we plot the upper and lower bound formulas for $E[cbtime]$ that we derived in Sec. 4. The blue lines indicate the bounds of Corollary 4.3, while the green line shows the weaker bound of Corollary 4.1 and the red line shows the trivial diameter bound. The jittered blue circles indicate the strongest bound for each spanning tree, derived from Observation 4.2. We observe that the lower bound is reasonably tight whereas the upper bounds based on Theorem 4.1 and average eccentricity are looser, but much better than the trivial diameter upper bound. The upper bound obtained from Observation 4.2 is tighter, since it takes the local structure of the spanning tree into account, rather than simply the globally maximum outdegree (46 in this case). In summary, this validates our analytical results.

6. CONCLUSION AND DISCUSSION

In this paper, we presented an analytical modeling framework (both models and metrics) for studying the composite stretch (or elongation) suffered by information on a multi-hop communication network when it is constrained by social network relationships. We derive scaling laws (in expected value sense) for composite stretch for random embeddings of common social network structures such as trees on a variety of multi-hop communication network models, ranging from simple linear networks to random geometric graphs.

We also analyze the constrained broadcast time metric, which measures the time taken to broadcast (or gossip) information along the edges of a social network to all nodes in that network while being constrained by the underlying
communication network structure. We derived analytical bounds for composite broadcast time and validated them by simulations using a friendship social network data set. We also show, using simulations based on a historical military deployment data set how the stretch of an information flow in a chain-of-command network is both non-optimal, but far superior to a random deployment scenario.

This is the first step toward an ambitious research program directed toward studying composite networks. Other composite metrics of interest include “composite load,” which measures the number of times a piece of information needs to traverse a particular edge in $G_2$. Additionally, other interesting variants of the constrained stretch metric exist – suppose one is only allowed to direct communication links that exist between friends (since these are supposed to be trusted) – this will result in a routing which is likely to have a higher stretch since shorter communication paths between friends (through non-friends) are disallowed. How can one characterize this highly constrained stretch? Such insights should help us design better communication networks (or facilitate intelligent deployment) that are suited to particular demands of the overlaid social networks of users.

7. REFERENCES