Layers and Hierarchies in Real Virtual Networks

Olga Goussevskaia
Computer Engineering and Networks Laboratory
ETH Zurich
8092 Zurich
golga@tik.ee.ethz.ch

Michael Kuhn
Computer Engineering and Networks Laboratory
ETH Zurich
8092 Zurich
kuhnmi@tik.ee.ethz.ch

Roger Wattenhofer
Computer Engineering and Networks Laboratory
ETH Zurich
8092 Zurich
wattenhofer@tik.ee.ethz.ch

Abstract

The virtual world is comprised of data items related to each other in a variety of contexts. Often such relations can be represented as graphs that evolve over time. Examples include social networks, co-authorship graphs, and the world-wide-web. Attempts to model these graphs have introduced the notions of hierarchies and layers, which correspond to taxonomies of the underlying objects, and reasons for object relations, respectively. In this paper we explore these concepts in the process of mining such naturally-grown networks. Based on two sample graphs, we present some evidence that the current models well fit real world networks and provide concrete applications of these findings. In particular, we show how hierarchies can be used for greedy routing and how separation of layers can be used as a preprocessing step to implement a location estimation application.

1. Introduction

Naturally-grown networks can be used to model and analyze phenomena in the physical world, in the virtual world, and in society. Therefore, the study of naturally-grown networks has received a great deal of interest in different communities. Social scientists analyze complex structures of social networks arising in the context of organizations. Biologists study the interactions of cells in intricate metabolic processes. Computer scientists face the emergence of gigantic online systems, ranging from online social networks, such as LiveJournal, to online collaborative data repositories, such as Wikipedia, not mentioning the Internet and the world-wide-web themselves.

After Milgram's famous "Six Degrees of Separation" experiment, many models have been proposed in the attempt to capture the discovered Small World property. The first models were mainly based on random graphs [2]. Soon, however, people understood that random graphs can only explain the short path lengths, but fail to reflect the high local clustering, and the fact that short paths can be found even if only local knowledge of the graph is available for routing. Ever better models appeared and finally the notion of social dimensions and hierarchies has been introduced. A social dimension thereby refers to a reason why one person chooses another person as a friend in, for example, the context of a social networking service. A hierarchy, on the other hand, refers to a taxonomy of these different reasons, and the models state that an edge becomes more likely the closer two persons are in this hierarchy structure.

In this paper we verify this conjecture and show that it also holds in non-social settings. Further, we generalize the notion of social dimensions and introduce the concept of *layers*. A *layer* basically is a subgraph that contains all the edges generated due to one particular reason. We will provide evidence for such layers and see that it is useful to understand the reasons behind object relations prior to applying techniques like classification, or clustering. The more diffuse these reasons are in a particular data set, the more difficult it becomes to develop techniques to automatically analyze the structure of the data. We therefore propose to separate the different layers of a graph and provide an approach to achieve such a separation if some information about the layered structure is known.

We validate the two model features (layers and hierarchies) using two different real world networks: the *Simple English Wikipedia* and the *LiveJournal* online social network. For *Wikipedia*, we show that the average path length between articles is related to the height of their least common ancestor in the category tree[1]. Moreover, we show that the category tree can be used to greedily find short paths in the graph using only information about the least common ancestor of the destination, the source node, and its direct

[1] Articles in *Wikipedia* can be classified according to a hierarchical set of categories, please refer to Section 4.

neighbors. For the *LiveJournal* social network, we show that the average geographical distance between friends[2] is related to their number of common interests. Surprisingly enough, the higher the number of common interests, the higher is the average geographical distance between friends. We use this fact to improve the precision of location estimation of nodes that lack coordinate information.

The rest of the paper is organized as follows. In Section 2 we discuss some related work. In Section 3 we describe the model upon which our experiments are based. In Section 4 we show that *Wikipedia* fits into the hierarchical model and discuss some applications. In Section 5 we show how the layered model can be used to improve location estimation in *LiveJournal*. We present our conclusions in Section 6.

2. Related Work

A variety of models has been proposed to describe the structure of naturally-grown graphs. The models most related to the techniques analyzed in this paper are the hierarchical and layered models, proposed by Kleinberg [6] and Watts [16] to analyze decentralized search in natural graphs.

The main idea is that individuals cluster the world hierarchically into categories. The deeper a category is in such a hierarchy, the more specific it is. The models further state that the social world can be clustered in more than one way (e.g. by geography and by occupation). Thereby each of these *social dimensions* is represented by an independently partitioned hierarchy, or layer. A node's identity is then defined as a multi-dimensional coordinate vector, in which each coordinate represents its position in each layer. In [16], a measure of similarity is defined as the minimum ultra-metric distance over all dimensions between two nodes. The intuition is that closeness in only one dimension is enough to connotate affiliation. A consequence of this metric is that social distance violates the *triangular inequality*.

More recently, another interesting model, focused on geographical properties of links, has been proposed in [4]. Using the data of *LiveJournal* as an experimental bed, it is argued that solely the distance between people is insufficient to explain the nature of friendships in a real social network, such as *LiveJournal*. The authors propose a model, where the probability that a person befriends a particular candidate is inversely proportional to the number of closer candidates. It is shown that such *density awareness* can be explored to discover short paths under geographic routing.

In the last couple of years, the structure of a number of emerging online networks has been analyzed from different perspectives, such as: evolution and dynamics [1, 5, 11], clustering [3, 12], embedding [8, 13], similarity and proximity analysis [7], among many others. The idea of preprocessing the graph and filtering links according to the "reason" behind them, has yet not been explored to improve any of these techniques.

[2]Members of *LiveJournal* are linked to other members through "friendship" links, and also can declare in their profiles to participate in areas of interest, please refer to Section 5.

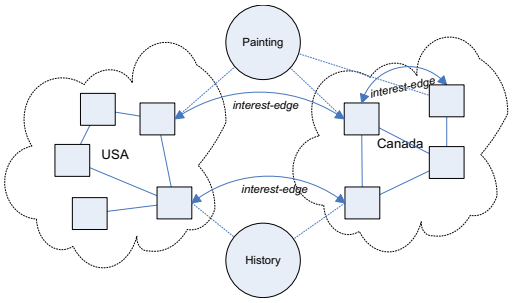

Figure 1. *LiveJournal*: **edges belong to different layers (geography, interests).**

Wikipedia has received a good deal of attention. Some interesting applications include automatic methods to compute semantic relatedness of words [8, 15], topic identification [14], and thesaurus construction [10]. The structure and evolution of *LiveJournal* has also shown to be an interesting topic. Characteristics such as main countries of origin, distributions of number of friends, age, language, bursts of activity over time, community membership, and nature of friendships have been discussed. An interesting fact, presented in [9], is that 70% of friendships can be "explained" by geographic location and interests of the users. We will come back to these links later.

3 Model

Throughout this paper we will assume that a naturally grown network exhibits two features:

- *Layers:* A layer is a subgraph containing all the edges that can be explained by a certain reason.
- *Hierarchies:* Each node in a graph belongs to one or more categories. These categories are hierarchically organized and can be represented by a tree.

The two concepts are illustrated in Figure 1 (layers) and Figure 2 (hierarchies). Note that these concepts are not independent. Rather, the semantic layers can often be hierarchically classified into groups of increasing granularity. Therefore, two people, both interested in glass painting, for example, are more likely to get acquainted, due to their hobby (or profession), than a person interested in glass and another person interested in fresco painting.

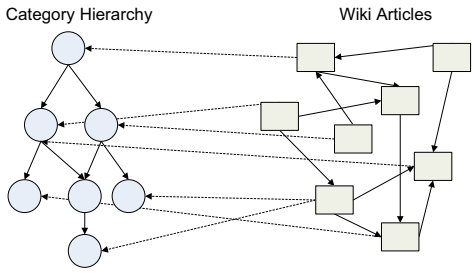

Figure 2. *Wikipedia*: **The articles are classified according to categories in the category DAG.**

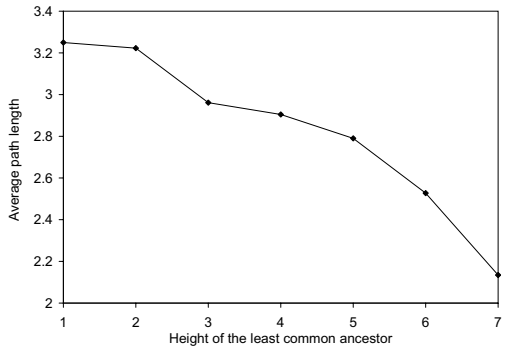

Figure 3. The deeper in the category DAG is the LCA, the shorter is the average path length between two articles.

4 Wikipedia

In this section we show that the *Simple English Wikipedia*[3], which we refer to just as *Wikipedia*, can be modeled using the before-mentioned hierarchical approach. We then take advantage of this representation, showing that greedy routing can be applied to efficiently find short paths in the graph.

The *article graph* of *Wikipedia* consists of articles (nodes) and the hyperlinks between articles (edges). In addition to the direct linkage of articles, *Wikipedia* contains a structured organization of topics, the *category tree* (Figure 2). As already mentioned in Section 3, the category tree is a hierarchical representation of topics that subdivides coarse grained general terms (such as *History*) into ever finer grained terms (such as *History of America* or *History of Oceania*) and finally reaches very specialized categories (such as *Physicians in the American Revolution*, or *Political leaders of the American Revolution*). In fact, the category tree is rather a directed acyclic graph (DAG) than a tree in the case of *Wikipedia*, as nodes can contain more than one parent.

We define the similarity $sim(C_i, C_j)$ between two categories C_i and C_j to be the height of their *Least Common Ancestor* (LCA) in the category DAG:[4]

$$sim(C_i, C_j) = height(LCA_{DAG}(C_i, C_j)). \quad (1)$$

Further, we define $C(A_k)$ to be the set of all categories which an article A_k belongs to. Thereafter, we define the similarity $sim(A_k, A_l)$ between two articles to be the maximum similarity among all pairs of categories to which articles A_k and A_l belong:

$$sim(A_k, A_l) = \max_{C_i \in C(A_k), C_j \in C(A_l)} (sim(C_i, C_j)) \quad (2)$$

[3] `http://simple.wikipedia.org`

[4] We assume that the root of the DAG has height zero.

4.1 Evidence for the Hierarchy-Model

The category graph of *Wikipedia* closely resembles the hierarchical structures Kleinberg [6] and Watts et. al [16] introduced in their small-world models. We thus decided to verify whether the models well describe the relations between the *Wikipedia article graph* and the *category DAG*, both naturally grown structures taken from the real world.

If the models are indeed correct, then nodes that are "closer" in the tree should exhibit a higher connectivity, and consequently the path length between such nodes should become shorter. We therefore constructed an experiment that compares the average path length between articles to the height of their least common ancestor (LCA) in the category DAG. In Figure 3, we can observe that the higher the similarity between two articles, i.e., the deeper in the category DAG their LCA, the shorter the average path between them. This confirms the expectations and shows that the model resembles the real world structure.

The average path length for each value of LCA was calculated upon a random sample of 20K pairs of articles. Comparing the approximate mean path length value of 3.163 with a variance of 0.645 to the path lengths measured in the experiments shows that the deviations are not random. Even the smallest deviation (for LCA = 2), has a probability of only about 0.6% to occur for a random set of pairs. This analysis shows that the probability of an edge between two nodes is indeed proportional to their semantic relatedness, therefore validating our hypothesis.

4.2 Greedy Routing

We apply the fact that the category tree influences path lengths between articles, as demonstrated in the previous section, to implement greedy routing in *Wikipedia*. Figure 3 suggests that the distance in the category graph can be used

for greedy routing, since short distances in the category graph indicate short distances in the article link structure. Our experiment consists in implementing a simple routing algorithm (see Algorithm 1) that uses the category graph to find short paths in the article graph. The only data that a node requires to forward a message is the similarity measure $sim(A_k, A_d)$, defined in (2), between each of its neighbors A_k and the destination article A_d. The greedy step consists in forwarding the message to the neighbor with highest similarity measure to the destination. This process continues until an article in the closest category to the destination is reached. Since the greedy step cannot go beyond the granularity of the category graph, once the closest category is reached, a normal flooding is applied to find the destination article among all articles that belong to that category.

Algorithm 1 Greedy Algorithm for *Wikipedia*

Require: Category DAG, destination article A_d, set of articles P visited until this point;

1: $d_{min} = \max_{C_i \in A_d}(height(C_i))$;
2: *Choose neighbor* A_k *with maximum* $sim(A_k, A_d)$;
3: **if** $(A_k = A_d)$ **then**
4: *Halt*;
5: **end if**
6: **if** $(A_k \in P)$ **then**
7: $A_k =$ *random neighbor*; (to avoid cycles)
8: **end if**
9: **if** $(sim(A_k, A_d) < d_{min})$ **then**
10: *Forward message to* A_k;
11: **else**
12: *Start Flooding*; $(sim(A_k, A_d) = d_{min}) \Rightarrow$ arrived at the lowest category subtree of A_d and greedy cannot make the distance shorter anymore.
13: **end if**

We compare the performance of the greedy algorithm to a Dijkstra-based alternative. Note that the algorithm of Dijkstra for finding shortest paths is basically the same as flooding in the case of unweighted graphs. The evaluation of the greedy routing focuses on the following three attributes:

- *Path stretch*: The factor by which the greedy-path gets longer than the shortest path.
- *Flooding stretch*: The factor by which the flooding part of the path becomes shorter.
- *Number of nodes visited*: The relation of the number of nodes visited using flooding and the number of nodes visited using greedy routing (including the final flooding part). Nodes that are visited multiple times are thereby counted multiple times.

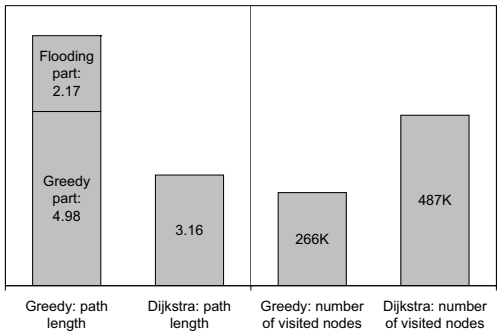

Figure 4. Greedy routing in Wikipedia.

In Figure 4 it is shown that the average path length obtained by the greedy routing was 7.15 hops, whereas the average shortest path was 3.16. This low *path stretch* indicates that it is indeed possible to find short paths with such an approach. Analyzing further the results in Figure 4, we can see that, on average, greedy forwarding performed 4.98 steps, and the remaining 2.17 hops were achieved by flooding. Compared to the length of the average shortest path, this gives a *flooding stretch* of 0.69, which results in 45% less nodes that need to be explored in order to reach the destination.

These values show that greedy routing can be an interesting option in graphs for which a hierarchical structure such as the *Wikipedia's* category DAG is known. The average path length increases only by a small constant. The overhead of visiting a huge number of nodes in the graph using flooding algorithms can, on the other hand, be drastically reduced.

For greedy routing to become applicable, it is important to be able to quickly calculate the (category DAG-) distance between two nodes. For our experiments we decided to pre-compute all the pairwise distances (i.e. LCAs) between categories. However, if the number of categories becomes large this approach becomes infeasible, since the corresponding table grows with $O(n^2)$. For larger category hierarchies, graph labeling provides an elegant alternative. An intelligent label l_i is attached to each category c_i. The labels are chosen such that they are short and that a distance function $d(l_i, l_j)$ can quickly be evaluated.

5 Live Journal

In this section we show that an online social network, such as *LiveJournal*, can be modeled using the layered approach described in Section 3. Furthermore, we show that if one of the layers is filtered from the data as a preprocessing step, a better location estimation can be obtained on the remaining graph.

Crawled nodes	250K
US nodes w/ geo. info.	88K
US edges	202K
Interests	673K

Table 1. Input data sample.

LiveJournal is a popular social networking site that currently counts about 13 million users.[5] As usual for social networking services, these users are connected to each other by friendship relations and thereby form a social graph. Further, *LiveJournal* users can create and participate in interest groups, and they can indicate their place of residence. As opposed to many similar services, *LiveJournal* is freely crawlable, which allows to retrieve and study aforementioned friendship relations and interest memberships.

Clearly, a major catalyst for friendships is geographic proximity. However, there must also exist other catalysts that create long-range contacts, which are the reason for the small-world character of social graphs. As in real-world, common interests are likely to cause friendship links also in the virtual world. Therefore, it is natural to expect that besides the geography layer, a second important layer in *LiveJournal* would be the *interest layer* (see Figure 1).

The attributes of the input data sample that we were able to crawl and use in our experiments are summarized in Table 1. Note that from 250K crawled users, only about 190K have indicated their country of residence, being 88K of them from the US. We will refer to the subset of US residents with known country and city of residence (i.e. known coordinates) as *US subset*.

5.1 Evidence for the Layer-Model

Assuming the layer model holds in the case of *LiveJournal*, we can conjecture that links in the geographic layer are typically shorter than links in the interest layer. Links that purely belong to the interest layer should have random length. Links in the geographic layer, on the other hand, should show a clear trend toward short links, as the corresponding friendships are supposed to be caused by geographic proximity.

Using the interest information in *LiveJournal* together with the geographic coordinates in the *US subset*, it is possible to verify the above conjecture. One distance unit in our experiment corresponds to one degree in the longitude and latitude space. For simplicity we assumed that this space is an Euclidean plane for the area of the United States.

We define a *geo-edge* to be a friendship link between two users that do not share any interests. An *interest-edge*, on the other hand, is a friendship link between users that share at least one common interest but that does *not* "close" a triangle with two *geo-edges*, i.e., two users linked by an *interest-edge* do not have any common friends in their physical proximity. If they have such a friend, the link is supposed to lie in the intersection region between the *geography* and the *interest layer*. In our analysis, we do not take such "intersection" links into account.

[5]Source: *Wikipedia*

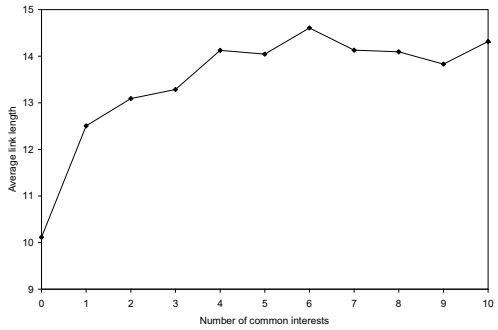

Figure 5. *LiveJournal*: **The average geographic length of a friendship link drastically increases when there is a common interest.**

For each edge in the friendship graph, we counted the number of common interests of the two corresponding users. It can be seen in Figure 5 that the average length of a friendship link increases as the number of common interests increases until a threshold of approximately 5 and then stabilizes. Given that *interest-edges* present an almost 20% increase in average geographic length, we can deduce that such links are less influenced by geographic proximity of residence and can be put into the *interest layer* of the friendship graph. To measure the statistical significance of our experiment, we performed 1000 random experiments, removing the same number of random edges as the number of *interest-edges*. The corresponding mean and variance values show that our identified *interest-edges* were not random, as the observed deviations from the mean value would occur with a probability of less than 0.3% in a random setting. This fact confirms our conjecture and thereby supports the layer model.

5.2 Location Estimation

In this section we propose an application for the properties exposed above. We implemented a simple location estimation application to prove the concept. The experiment was performed on the *US sample*. We randomly selected 50% of the nodes to serve as landmarks, and estimated the location of the other 50% of the nodes. The estimated position was set to be the center of mass of its adjacent landmarks (see Figure 6). The location estimation was

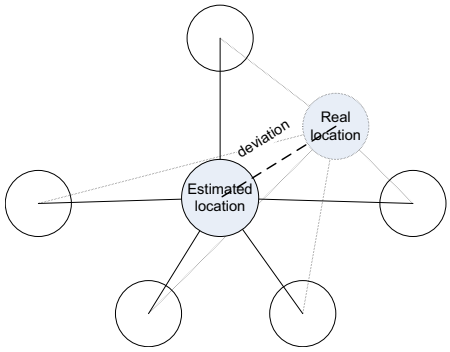

Figure 6. *LiveJournal*: **The estimated position of a node is the center of mass of its friends.**

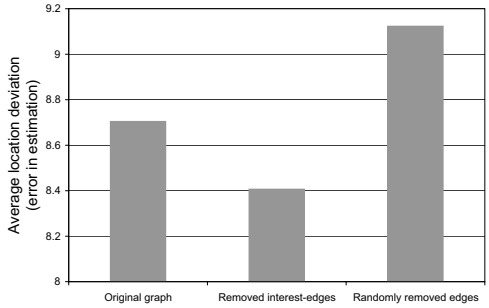

Figure 7. *LiveJournal*: **Location estimation is more accurate if only** *geo-edges* **are used.**

performed on the original friendship graph and compared to a preprocessed graph, with all *interest-edges* removed from it. Our results on 1000 such experiments showed that the preprocessed graph resulted in an average 8.4-unit location deviation, as opposed to 8.7-unit deviation in the original "mixed layer" graph. To measure the statistical significance of our experiment, we performed 1000 random experiments, removing the same number of random edges as the number of *interest-edges*, which resulted in a 9.1-unit average deviation, giving a significance level of approximately 0.95% (see Figure 7).

6 Conclusions

In this paper we intended to throw some insights into how layered and hierarchical models for small world graphs could be applied in data mining techniques. We present some evidence that such models can be used to characterize two different online natural networks and discuss some applications. We believe that, in applications where information about how to identify one or more layers is available, the separation of layers as a preprocessing step can improve the performance of many data mining techniques. How such layer separation can be generalized and function independently of a particular data set, however, remains to be investigated in future work.

References

[1] L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54, 2006.

[2] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev.*, August 2004.

[4] R. K. P. R. David Liben-Nowell, Jasmine Novak and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–1162, 2005.

[5] T. Falkowski, J. Bartelheimer, and M. Spiliopoulou. Mining and visualizing the evolution of subgroups in social networks. In *Web Intelligence*, pages 52–58, 2006.

[6] J. Kleinberg. Small-World Phenomena and the Dynamics of Information. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[7] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *KDD*, pages 245–255, 2006.

[8] M. Kuhn and R. Wattenhofer. Community-Aware Mobile Networking. In *1st Workshop on Mobile Services and Personalized Environments (MSPE), Aachen, Germany*, November 2006.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, 2004.

[10] D. Milne, O. Medelyan, and I. H. Witten. Mining domain-specific thesauri from wikipedia: A case study. In *ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pages 442–448, 2006.

[11] J. Novak, R. Kumar, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the Twelth ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (poster)*, 2006.

[12] J. C. Paolillo, S. Mercure, and E. Wright. The social semantics of livejournal foaf: Structure and change from 2004 to 2005. In *Proceedings of the 1st Workshop on Semantic Network Analysis at the ISWC 2005 Conference*, pages 69 – 80, Galway, Ireland, November 2005.

[13] J. C. Platt. Fast embedding of sparse music similarity graphs. *Advances in Neural Information Processing Systems*, 16:571578, 2004.

[14] P. Schonhofen. Identifying document topics using the wikipedia category network. *WI*, 0:456–462, 2006.

[15] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. July 2006.

[16] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and Search in Social Networks. *Science*, 296:1302–1305, 2002.