Replaying the Geometric Growth of Complex Networks and Application to the AS Internet

Fragkiskos Papadopoulos Constantinos Psomas
Department of Electrical Engineering, Computer Engineering and Informatics
Cyprus University of Technology
{f.papadopoulos, c.psomas}@cut.ac.cy

Dmitri Krioukov
Cooperative Association for
Internet Data Analysis
dima@caida.org

1. INTRODUCTION

Our growing dependence on networks has inspired a burst of research activity in the field of network science. One focus of this research is to derive network models capable of explaining common structural characteristics of large real networks, such as the Internet, social networks, and many other complex networks [1]. A particular goal is to understand how these characteristics affect the various processes that run on top of these networks, such as routing, information sharing, data distribution, searching, and epidemics [1]. Understanding the mechanisms that shape the structure and drive the evolution of real networks can also have important applications in designing more efficient recommender and collaborative filtering systems [2], and for predicting missing and future links—an important problem in many disciplines [3].

Krioukov et al. [4] have shown that there are intrinsic connections between complex network topologies and hyperbolic geometry, since the former exhibit hierarchical, tree-like organization, while the latter is the geometry of trees [5]. Following [4], Papadopoulos et al. [6], have recently shown that trade-offs between popularity and similarity shape the structure and dynamics of growing complex networks, and that these trade-offs in network dynamics give rise to hyperbolic geometry. The work in [6] introduces a simple model for constructing synthetic growing networks in the hyperbolic plane, which *simultaneously* exhibit many common structural and dynamical characteristics of real networks. We call the model of [6] the Popularity×Similarity Optimization (PSO) model.

Given the ability of the PSO model to construct synthetic growing networks that resemble real networks across a wide range of structural and dynamical characteristics, an interesting question is whether one can reverse the synthesis, and given a real network, map (embed) the network into the hyperbolic plane, in a way congruent with the PSO model. Our main contribution in this work is an affirmative answer to this question and a systematic framework that accomplishes this task, by replaying the network's geometric growth. The proposed framework, called *HyperMap*, is quite simple and it is supported by theoretical analysis. We apply this framework to the Autonomous Systems (AS) topology of the real Internet and show that it produces meaningful results, identifying communities of ASs that belong to the same geographic region. Further, we show that the proposed framework has a remarkable predictive power, demonstrated by its ability to predict missing links with high precision. While we consider here the AS Internet topology and the prediction of missing links [3], there are also other interesting areas where the proposed framework could find applications, e.g., in community detection [1], and in the prediction of future links [3].

2. PRELIMINARIES: THE PSO MODEL

The PSO model [6] constructs a growing network up to $t > 0$ nodes as follows: (1) initially the network is empty; (2) at time $1 \leq i \leq t$, new node i appears having coordinates (r_i, θ_i), where $r_i = 2 \ln i$, while θ_i is uniformly distributed on $[0, 2\pi]$, and every existing node j, $j < i$, moves increasing its radial coordinate according to $r_j(i) = \beta r_j + (1-\beta) r_i$ with parameter $\beta \in [0, 1]$; and (3) node i looks at every existing node j, $j < i$, and connects to it with probability $p(x_{ji}) = \frac{1}{1+e^{\frac{1}{2T}(x_{ji}-R_i)}}$, where x_{ji} is the hyperbolic distance between nodes j and i, $\cosh x_{ji} = \cosh r_j \cosh r_i - \sinh r_j \sinh r_i \cos\theta_{ji}$, with $\theta_{ji} = \pi - |\pi - |\theta_j - \theta_i||$, and $R_i = r_i - 2\ln[\frac{2T}{\sin T\pi} \frac{1-e^{-\frac{1}{2}(1-\beta)r_i}}{m(1-\beta)}]$. Model parameter m is the average number of existing nodes that a new node connects to, defining the average node degree in the network $\bar{k} = 2m$. Parameter $\beta \in [0, 1]$ is a function of the exponent $\gamma \geq 2$ of the target power law degree distribution $P(k) \sim k^{-\gamma}$, $\beta = \frac{1}{\gamma-1}$. Finally, model parameter $T \in [0, 1)$ is called temperature, and controls the average clustering in the network: clustering is maximized at $T = 0$, and it decreases to zero as $T \to 1$. To construct a network up to t nodes we need to specify m, β, and T.

It has been proven in [6] that the expected degree $\overline{k_i(t)}$ of a node born at time i by time $t \geq i$ is $\overline{k_i(t)} \sim (\frac{t}{i})^{\frac{1}{\gamma-1}}$. This equation says that the earlier a node appears the higher is its expected degree. We use this observation in HyperMap in the next section. In Figure 2(a) we use real data [7] to validate that this equation indeed describes the trend in the evolution of the average degree of an AS in the Internet as a function of the time the AS appeared. To draw Figure 2(a) we took the historical data of the twelve-year (1998-2010) evolution of the AS Internet from [7], and for each AS we found the time i (number of nodes present in the network) when the AS first appeared in the data. Then, for all ASs that appeared at time i, and which were present at the end of the measurement period (where $t = 33796$ nodes) we calculated their average degree $\overline{k_i(t)}$. For the theoretical formula we used $\gamma = 2.1$, i.e., the γ of the AS Internet [1].

The PSO model reproduces not only the degree distribution and clustering of real networks, but also many other important properties [6]. Given the ability of the PSO model to construct growing synthetic networks that resemble real networks, we show that it is possible to reverse the synthesis, and given a real network, to map (embed) the network into the hyperbolic plane, in a way congruent with the PSO model.

3. THE MAPPING METHOD (HYPERMAP)

Given a scale-free network with t nodes, average node degree $\bar{k}$, power law exponent $\gamma \geq 2$, temperature $T \in [0, 1)$, and adjacency matrix $\{\alpha_{ij}\}$—$\alpha_{ij} = \alpha_{ji} = 1$ if there is a link between nodes i and j, and $\alpha_{ij} = \alpha_{ji} = 0$ otherwise—HyperMap computes radial and angular coordinates $r_i(t), \theta_i$, for all nodes $i \leq t$ as shown in

1: Sort node degrees in decreasing order $k_1(t) > k_2(t) > ... > k_t(t)$ with ties broken arbitrarily.
2: Call node i, $i = 1, 2, ..., t$, the node with degree $k_i(t)$.
3: Node $i = 1$ is born, assign to it initial radial coordinate $r_1 = 0$ and random angular coordinate $\theta_1 \in [0, 2\pi]$.
4: **for** $i = 2$ to t **do**
5: Node i is born, assign to it initial radial coordinate $r_i = 2 \ln i$.
6: Increase the radial coordinate of every existing node $j < i$ according to $r_j(i) = \beta r_j + (1 - \beta) r_i$.
7: Assign to node i angular coordinate θ_i maximizing L_i given by Eq. (1).
8: **end for**

Figure 1: The HyperMap Embedding Algorithm.

Figure 1.

Specifically, HyperMap first estimates the order by which the nodes of the network are born. Since, according to the PSO model, the earlier a node appears the higher its expected degree, HyperMap first computes the degree of every node in the network and then sorts the node degrees in the decreasing order $k_1(t) > k_2(t) > \ldots > k_i(t) > \ldots > k_t(t)$, with ties broken arbitrarily, thus creating a sequence of node birth times $i = 1, 2, \ldots, t$, corresponding to nodes with degrees $k_1(t), k_2(t), \ldots, k_i(t), \ldots, k_t(t)$. We call the node born at time i node i. Having a sequence of node birth times, HyperMap replays the geometric growth of the network in accordance with the PSO model as follows. When a node is born at time $1 \leq i \leq t$, it is assigned an initial radial coordinate $r_i = 2 \ln i$, and every existing node $j < i$ moves increasing its radial coordinate according to $r_j(i) = \beta r_j + (1 - \beta) r_i$, with $\beta = \frac{1}{\gamma - 1}$. To compute the angular coordinate θ_i of a new node i, we first define *likelihood* L_i:

$$L_i = \prod_{1 \leq j < i} p(x_{ji})^{\alpha_{ji}} \left[1 - p(x_{ji})\right]^{1 - \alpha_{ji}}, \quad (1)$$

where x_{ji} is the hyperbolic distance between node i and existing node j, $p(x_{ji})$ is the connection probability defined in the previous section, and α_{ji} is the network adjacency matrix. Likelihood L_i is the probability that the *given* set of connections between new node i and existing nodes $j < i$ take place in the PSO model. This likelihood is a function of θ_i, since x_{ji} depends on θ_i, $p(x_{ji})$ depends on x_{ji}, and L_i depends on $p(x_{ji})$. The best value for θ_i is then the value that maximizes L_i. The maximization can be performed numerically, by sampling the likelihood L_i at different values of θ in $[0, 2\pi]$ separated by intervals $\Delta\theta = O(\frac{1}{i})$, and then setting θ_i to the value of θ that yields the largest value of L_i. Since to compute L_i for a given θ we need to compute the connection probability between node i and all existing nodes $j < i$, we need a total of $O(i^2)$ steps to perform the maximization. We note that since L_i is sampled at θ values separated by $O(\frac{1}{i})$ intervals, the maximization is approximate and becomes more precise as i increases.

Since the PSO model can construct growing synthetic networks that resemble real networks, we expect HyperMap to be able to accurately map a given real network into the hyperbolic plane, in a way congruent with the PSO model. In the next section we use the AS Internet topology to show that this is indeed the case. We note that HyperMap uses the *current* network adjacency matrix in Equation (1) to find the best estimate for the angular position of each node, and does not require any knowledge about whether nodes/links were departing while the network was evolving, or whether connections between some nodes might have been internal (i.e., took place some time after the nodes appeared). More details related to these remarks and to the method will appear in a longer version of this paper.

4. VALIDATION

After mapping a network with t nodes, we have the radial and angular coordinates $r_i(t), \theta_i$, for all nodes $i \leq t$, and therefore, we can compute the hyperbolic distance between every pair of nodes. To evaluate how well HyperMap maps the network we use two metrics: (i) the connection probability $\tilde{p}(x(t))$, which is the probability that there is a link between a pair of nodes given their hyperbolic distance $x(t)$ at time t; and (ii) the distance distribution $d(x, t)$, which is the percentage of node pairs whose hyperbolic distance at time t is x. After mapping a network we compute these two metrics and juxtapose them against our theoretical predictions for networks growing according to the PSO model, given below:

$$\tilde{p}(x(t)) \approx \frac{1}{t - l_{min} + 1} \sum_{l = l_{min}}^{t} \frac{1}{1 + e^{\frac{1}{2T}(x(t) - 4(1-\beta)\ln\frac{t}{l} - R_l)}}, \quad (2)$$

where $l_{min} = \max\left(2, \lceil t e^{-\frac{x(t)}{4(1-\beta)}} \rceil\right)$, and

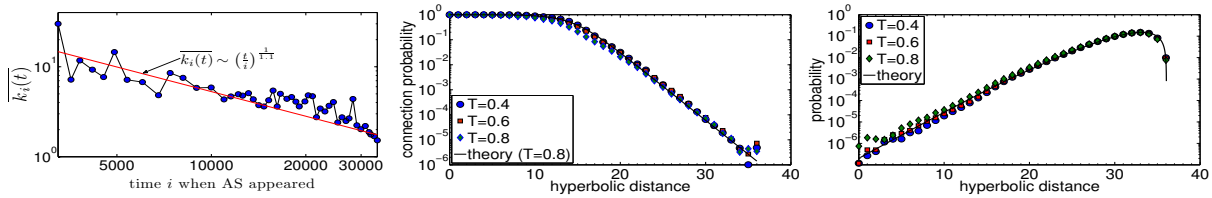
$$d(x,t) \approx \left(\frac{\pi(1-\beta)-2}{4\pi\beta^2(1-\beta)} x + \frac{1}{\pi(1-\beta)^2}\right) e^{\frac{1}{2\beta}(x - 2r_t)} + \frac{1}{\pi(1-\beta)^2}\left(e^{\frac{1}{2}(x-2r_t)} - 2e^{\frac{1}{2}(x-(1+\frac{1}{\beta})r_t)}\right), \quad (3)$$

if $x \leq r_t$, or otherwise,

$$d(x,t) \approx \left(\frac{\pi(1-\beta)-2}{4\pi\beta^2(1-\beta)}\right)(2r_t - x) e^{\frac{1}{2\beta}(x - 2r_t)} + \frac{1}{\pi(1-\beta)^2}\left(e^{\frac{1}{2}(x-2r_t)} - e^{\frac{1}{2\beta}(x-2r_t)}\right). \quad (4)$$

AS Internet. We use the AS Internet topology [8] of December 2009, available at [9]. The connections in this topology are not physical but logical, representing AS relationships [9]. We consider the topology consisting of all nodes (ASs) with degree greater than 2. There are $t = 8220$ such nodes, and the topology has a power law degree distribution with exponent $\gamma = 2.1$, average node degree $\bar{k} = 9.45$ and average clustering $\bar{c} = 0.60$. We map the topology using HyperMap with different values of the temperature T, and show the results in Figures 2(b),(c). From the figures we observe that HyperMap is remarkably accurate. Further, different T values give approximately the same results. In particular, all the T values give a connection probability that is best matched theoretically (using Eq. (2)) with $T = 0.8$. These results are interesting as they imply that in practice HyperMap is not very sensitive to the exact value of the input parameter T. As mentioned in Section 2, $\bar{c}$ is controlled by T. But given $\bar{c}$ and the network topology there is no formula that can be used to infer T. However, our results above suggest that we can find T for a real network experimentally, by embedding the network using different values for T, and then use Equation (2) to find the T value that best matches the empirical connection probability.

In Figure 3, we demonstrate that HyperMap produces meaningful results. The figure shows the angular distribution of ASs that belong to the same country, for 13 different countries. The AS-to-country mapping is taken from the CAIDA AS ranking project [10]. We observe that even though HyperMap is completely geography-agnostic, it discovers meaningful groups or communities of ASs belonging to the same country. The reason for this is that ASs belonging to the same country are usually connected more densely than the rest of the world, and HyperMap correctly places all such ASs in narrow regions, close to each other. However, as expected, due to significant geographic spread in ASs belonging to the US, these ASs are widespread in $[0°, 360°]$. We note that other reasons besides geographic proximity may affect the connectivity between ASs, such as



(a) AS degree evolution. (b) Connection probability. (c) Distance distribution.

Figure 2: AS degree evolution, connection probability, and distance distribution.

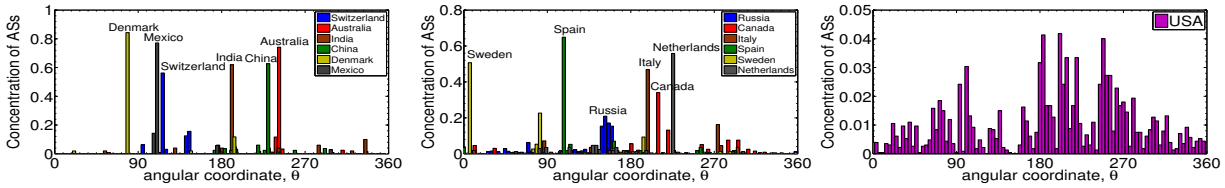


Figure 3: HyperMap yields meaningful results.

economical and/or political reasons. HyperMap does not favor any specific reason but relies only on the connectivity between ASs in order to place the ASs at the right angular (and consequently hyperbolic) distances.

5. APPLICATION TO THE PREDICTION OF MISSING LINKS

Topology measurements of many real networks, not only of the Internet [11], may miss some links. The prediction of missing links is a fundamental problem that attempts to estimate the likelihood of the existence of a missing link between two nodes in a network, based on the observed links and/or the attributes of nodes [3]. The standard way to evaluate a link prediction technique is to randomly remove a percentage of links from a given network topology, and then work with this incomplete data using the technique to see how well these missing links can be predicted. The standard metrics used to quantify the accuracy of a link prediction technique is the *Area Under the Receiver Operating Characteristic Curve (AUC)* and *Precision* [3]. A link prediction algorithm gives to each non-observed link (i, j) a score s_{ij} to quantify its existence likelihood. The prediction algorithm then orders all the non-observed links according to their scores, from the best score to the worst score. The AUC is the probability that a randomly chosen missing link is given a better score than a randomly chosen nonexistent link. If we consider only the top-L links from the ordered list, among which L_r links turn out to be right (i.e., indeed missing), then the Precision is the ratio $\frac{L_r}{L}$. Below, to compute Precision we use $L = 100$ (as used in [3]).

Performance of HyperMap. To check how effective HyperMap is in predicting missing links, we *first* remove 30% of links from the AS Internet topology and *then* embed the resulting topology using the method with $T = 0.8$. After the embedding, the score s_{ij} between a disconnected pair of nodes i, j, i.e., the score of each non-observed link (i, j), is the hyperbolic distance x_{ij} between the nodes i and j. The smaller this score, i.e., the smaller the hyperbolic distance between the two nodes, the more likely it is that a link between these two nodes is missing, since the connection probability $p(x_{ij})$ is a decreasing function of x_{ij}. Both AUC and Precision in HyperMap are remarkably high, AUC $= 0.95$, Precision $= 0.71$, indicating that the method has a strong predictive power.

6. FUTURE WORK

There are several directions for future work. One is to further explore and understand HyperMap's ability to predict missing links. Another, is to find efficient ways to expedite the running time of HyperMap without compromising the embedding accuracy. Finally, it would be interesting to explore the efficiency of HyperMap for other tasks, such as community detection (see Figure 3), or the challenging problem of predicting *future* links in different evolving networks.

7. REFERENCES

[1] S N Dorogovtsev. *Lectures on Complex Networks*. Oxford University Press, Oxford, 2010.

[2] A K Menon and C Elkan. Link Prediction via Matrix Factorization. In *ECML, LNCS 6912*, pages 437–452, 2011.

[3] L Lu and T Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390:1150–1170, 2011.

[4] D Krioukov, F Papadopoulos, M Kitsak, A Vahdat, and M Boguñá. Hyperbolic Geometry of Complex Networks. *Physical Review E*, 82:36106, 2010.

[5] M Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser, Boston, 2007.

[6] F Papadopoulos, M Kitsak, M Á Serrano, M Boguñá, and D Krioukov. Popularity versus Similarity in Growing Networks. June 2011. `http://arxiv.org/abs/1106.0286`.

[7] A Dhamdhere and C Dovrolis. Twelve years in the evolution of the Internet ecosystem. *IEEE/ACM Transactions on Networking*, 19(5):1420 –1433, oct. 2011.

[8] K Claffy, Y Hyun, K Keys, M Fomenkov, and D Krioukov. Internet Mapping: from Art to Science. In *CATCH*, pages 205–211. IEEE Computer Society, 2009.

[9] IPv4 Routed /24 AS Links Dataset. `http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml`.

[10] X Dimitropoulos, D Krioukov, M Fomenkov, B Huffaker, Y Hyun, K claffy, and G Riley. AS Relationships: Inference and Validation. *Comput Commun Rev*, 37(1):29–40, 2007.

[11] A Lakhina, J Byers, M Crovella, and P Xie. Sampling Biases in IP Topology Measurements. In *INFOCOM*, 2003.