Formatting Instructions for NIPS 2015

Anonymous Author(s)
Affiliation
Address
email

Abstract

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Experiments

1.1 Graph topology

GRID 2D lattices where each node has four neighbors except on the boundary. Synthetic ones are square while those labelled "real world" stem from the pictures showed in Figure 1.

PREFERENTIAL ATTACHMENT Synthetic ones are build according to [2], where new nodes are connected to 3.13 existing ones (that is either 3 or 4). These graphs are quite sparse and have short diameter, thus providing approximation of online social network. Real world PREFERENTIAL ATTACHMENT graphs include WIKIPEDIA, SLASHDOT and EPINION (from [5]) along with GOOGLE+ (Table 1). The last one was constructed from ego networks of GOOGLE+[1] and we kept the largest connected component of users of whom we know the gender.

TRIANGLE random 2D points triangulated using Delaunay algorithm.

Table 1: Dataset description

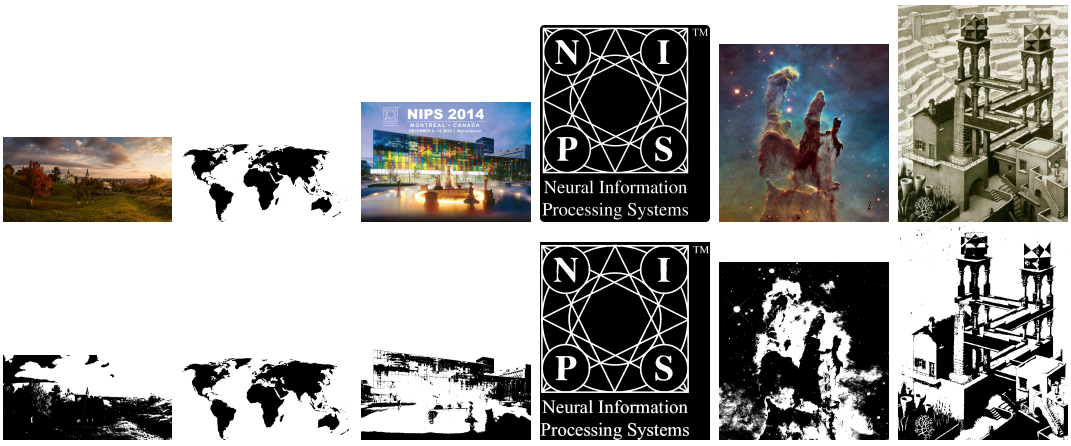
	$\lvert V\rvert$	$\lvert E\rvert$	fraction of $+$ edges	$\frac{2\lvert E\rvert}{\lvert V\rvert\cdot(\lvert V\rvert-1)}$
WIKIPEDIA	7065	99936	78.5%	$4.00 \cdot 10^3$
GOOGLE+	74917	10130461	67.6%	$3.61 \cdot 10^3$
SLASHDOT	82052	498527	76.4%	$1.48 \cdot 10^4$
EPINION	119070	701569	83.2%	$9.90 \cdot 10^5$

1.2 Stretch

The first property of Galaxy trees we wish to evaluate is their stretch, which depends only of graph topology. Namely, let G be a graph over vertex set V with $|V| = n$ and edge set E. Furthermore, let T be a spanning tree of G and E_{test} the edges of G not in T. Then we define the *average test edge stretch* as $\frac{1}{|E_{test}|} \sum_{(u,v)\in E_{test}} |path^T_{u,v}|$, where $path^T_{u,v}$ is the unique path between u and v in T.

As our graphs are unweighted, we compare with a spanning tree tree rooted at the highest degree node and obtained through a breadth first visit of the graph. This involve randomness as to which child

[1] Available at http://snap.stanford.edu/data/egonets-Gplus.html

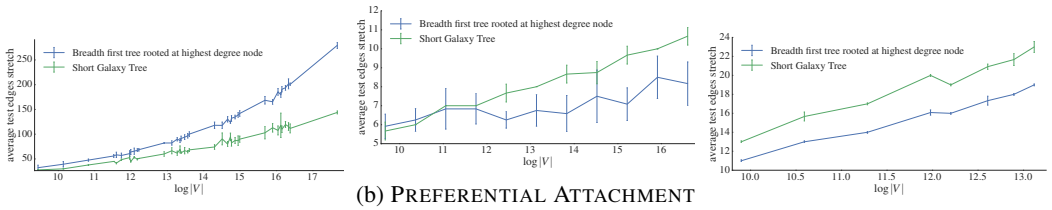


(a) monastery (b) world (c) poster (d) logo (e) space (f) waterfall

Figure 1: Real world pictures and their binarized version

node to choose next. Likewise in galaxy tree, the choice of the edge linking two stars is somewhat arbitrary. Therefore for each graph, we repeat the tree construction 12 times and present the average result, noting that the variance (showed as error bar in Figure 2) is small.

On PREFERENTIAL ATTACHMENT and TRIANGLE, we see that both trees exhibits logarithmic stretch, although with a larger constant in GALAXY TREE case. Note that this is the case for others low stretch tree methods [6, §5.3.1]. On GRID, GALAXY TREE maintain this logarithmic stretch growth while it is no longer the case for BREADTH FIRST TREE.



(a) GRID (b) PREFERENTIAL ATTACHMENT (c) TRIANGLE

Figure 2: Stretch over graphs of increasing size

1.3 Sign prediction

The second design goal of Galaxy trees is to accurately predict the sign of edges in E_{test}. Except for three real dataset that already include signs[2], all the other are constructed, meaning we have to set sign on their edges in the first place. This is done by partitioning the nodes into two clusters. For GOOGLE+ we use nodes gender, for pictures the node color and for all others, we propagate label from randomly selected high degree nodes.

We evaluate the performance of our prediction using the Matthews Correlation Coefficient (MCC)[1]

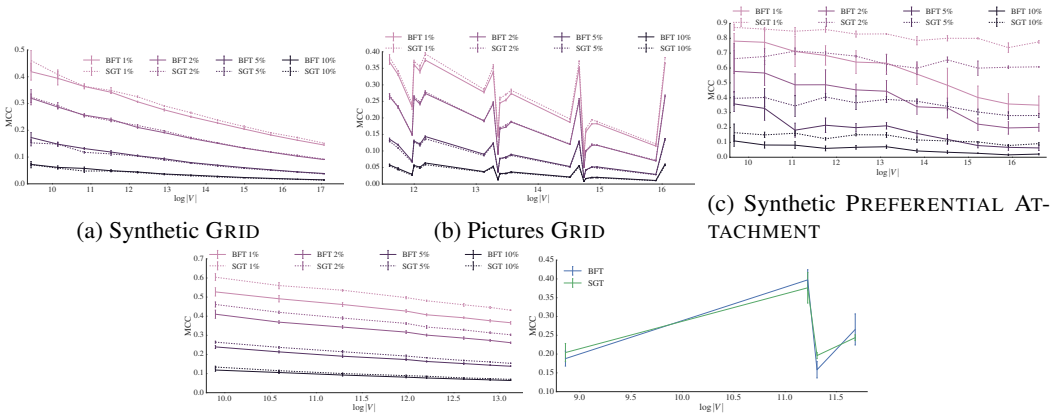
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} = \pm\sqrt{\frac{\chi^2}{n}}$$

Since we do not have confidence score, we cannot use AUC. Yet we have to account for the large sign unbalance and thus cannot rely on accuracy or F_1 measure. Therefore we choose MCC, which combine all the four numbers of the confusion matrix in a single metric. It ranges from $+1$ (perfect prediction) to -1 (inverse prediction) through 0 (random prediction). As a further motivation,

[2]We nonetheless perform some preprocessing as these 3 graphs are directed and include a small proportion of conflicting edges (e.g. positive from u to v but negative from v to u).

predicting all edges but one to be positive on Slashdot gives .764 accuracy, .886 F_1 score but -0.0007 MCC.

In most cases, we set the signed to be perfectly consistent with the clustering and predicting using path parity will thus give prefect result. To test performance in real or adversarial situation, we add noise, that is select of fraction of edges at random and reverse their sign. As we can see in Figure 3, when the noise level is low, GALAXY TREE performs better than BREADTH FIRST TREE. As it gets higher, they have similar performance. Note also than in Figure 3c, GALAXY TREE is less sensible to the size of the graph.



(a) Synthetic GRID (b) Pictures GRID (c) Synthetic PREFERENTIAL ATTACHMENT

(d) TRIANGLE (e) Real world network

Figure 3: MCC over various graphs

To further assess the quality of our trees, we plug them in them into a successful heuristic method to predict edge sign: *A sym exp*[4] It computes the exponential of the adjacency matrix after it has been reduce to z dimension. This allow to count the sign of all paths between two pairs of nodes with decreasing weight depending of their length. To simulate an active learning setting, we reveal only a subset of edge in A. This subset can be: $i)$ the edges forming a BREADTH FIRST TREE, $ii)$ the edges forming a GALAXY TREE $iii)$ $|V| - 1$ edges chosen uniformly at random.

We set the parameter z equal to 15 because $i)$ it is one of the best in [4, Fig. 11], $ii)$ it performs well on real dataset in [3, Fig.3], $iii)$ it was good in our initial testing (`20150401_wed_spectral.ipynb`)

As the *A sym exp* has a $O(n^3)$ complexity and uses quite some memory at prediction time, the larger graphs used previously are not all included. The conclusion of Figure 4 is that except on social network, it is better to use spanning trees than random edges. Specificaly, GALAXY TREE on GRID and BREADTH FIRST TREE elsewhere.

Finally we also compare GALAXY TREE with BREADTH FIRST TREE and RANDOM SPANNING TREE on the task of nodes prediction using Shazoo algorithm[7]. *Well I will start this week-end :)*

Acknowledgments

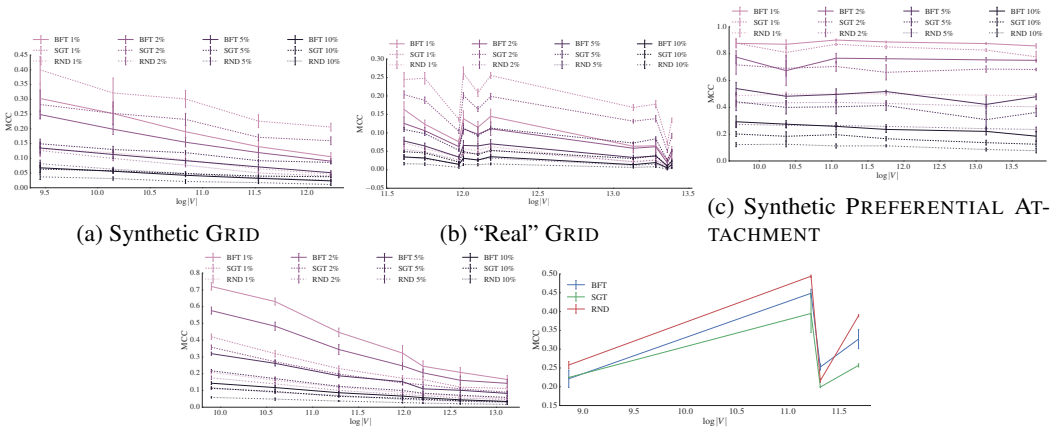
Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

[1] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview.," *Bioinformatics (Oxford, England)*, vol. 16, no. 5, pp. 412–424, 2000.

[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella, "A linear time active learning algorithm for link classification," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012, pp. 1–12.



(a) Synthetic GRID (b) "Real" GRID (c) Synthetic PREFERENTIAL ATTACHMENT

(d) TRIANGLE (e) Real world network

Figure 4: *A sym exp* over various graphs

[4] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The slashdot zoo: mining a social network with negative edges," in *Proceedings of the 18th international conference on World wide web - WWW '09*, 2009, p. 741.

[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010, p. 641.

[6] P. A. Papp, S. Kisfaludi-Bak, and Z. Király, "Low-stretch spanning trees," Bachelor thesis, Eötvös Loránd University, 2014.

[7] F. Vitale, N. Cesa-Bianchi, C. Gentile, and G. Zappella, "See the tree through the lines: the shazoo algorithm," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 1584–1592.