First year progress report

Géraud Le Falher

September 9, 2015

1 Scientific work

1.1 Context

Graphs are simple yet powerful abstractions to model relationships between entities. Hence, they have been used to represent social networks, linked data on the web, human cortex, scientific collaboration, physical networks like roads or power grid, protein interaction and so on. In many cases, these graphs adhere to the homophily assumption, meaning that nodes of the graph are connected because they share similar properties. In social context, this is known as "birds of a feather flock together" [1]. Machine Learning algorithms rely on this assumption to perform classification.

However, already in the fifties, sociologists have extended this model to take negative relationships into account [2, 3], namely link expressing the dissimilarity between two nodes. In social network, this can be viewed as distrust or dislike between two users. More broadly, one protein may hinder the action of another, and the rise of a city popularity may negatively affect its neighbors.

Such graphs are called signed graphs and their impact on graph learning is the subject of my thesis. More specifically, we identified several Machine Learning graph tasks that need to be tailored to this extended model. These include clustering, which in this context is named Correlation Clustering (CC) and classification, whether of nodes or edges (called in that case Link Classification).

1.2 Problems

In Correlation Clustering [4], given a signed graph as input, we want to find a partition of the nodes minimizing the number of *disagreement edges* (i.e. positive edges between clusters and negative edges within clusters). This is useful to perform entities resolution (merging the records from several databases that refer to the same instance using external similarity information), to find co-references in text, to aggregate several clusterings of the same data or to identify genes relationship.

In Link Classification [5], we are also given a signed graph, but some signs are hidden and we want to predict them, given the graph structure and the known signs of the other edges. Examples of application include understanding trust dynamic and communities formation, testing social theories at a large scale and recommending products from feedback in bipartite users/products graphs.

1.3 State of the art

1.3.1 Correlation Clustering

CC is a APX-hard (i.e. there exists $c > 1$ such that it is NP-hard to approximate with a ratio of c). There are several lines of research to go around this fundamental limitation:

- approximation algorithms, which are based on Linear or Semi-definite Programming (for complete or general graphs respectively). The best approximation ratios to date are $(2.06 - \epsilon)$ for complete graphs [6] and $O(\log n)$ for general graphs [7]. Also worth mentioning is a simple randomized combinatorial algorithm (called KwikCluster) with an expected ratio of 3 for complete graph [8]. On a practical side, a recent trend seems to be considering any graph as complete by assuming that its missing edges are negative.

- heuristic approaches, which do not come with theoretical guarantees yet provide good performances [9, 10].

As current graphs can be very large, scalability becomes an issue and requires new approaches. For instance parallel computation, which takes advantage of the multi cores of modern hardware architecture [11, 10]. When the graph is too large to fit in the memory of a single machine, it is possible to partition it and use MapReduce formalism to compute a solution in several rounds of message passing [12]. Yet another idea is to stream the edges of the graph [13].

Although CC formulation is slightly different, it is also related with two other well establish graph partitioning problems, which have been recently extended to handle negative links: community detection [14, 15, 16, 17] and spectral clustering [18, 19].

1.3.2 Link Classification

After the work of Leskovec *et al.* [5], who trained a logistic regression on triangle patterns of each edge, there have been more supervised approaches looking at higher order cycles [20], training SVM on graphlets (small subgraphs) [21], embedding the edges in a low dimensional space [22], or using transfer learning [23].

Departing from the batch setting, some works focus on the active scenario. There, the learner can first select some edges — whose signs will be revealed — before starting to make prediction [24, 25]. The goal is therefore to select as few edges as possible while minimizing the prediction error on the testing set.

1.4 First results

I started working on CC, first by writing a state of the art, which is summarized above. Then we studied how to transfer the combinatorial algorithm KwikCluster to general graphs while preserving the $O(\log n)$ approximation. The idea was to add missing edges with a sign that did not introduce bad triangles (i.e. triangle with a single negative edge, as such triangles always induce a disagreement edge no matter the clustering). This dependency on triangles proved to be costly on the running time, hurting the scalability. Furthermore, experimental results were mixed. Namely, performances strongly varied with respect to the order in which edges were added, in ways that we were not able to fully explained. Therefore, we concluded that this long open standing problem should be attacked first by focusing on interesting subclass of graphs that have yet to be identified.

This summer, I supervised Paul Dennetiere's internship in our team. He implemented the parallel version of KwikCluster described in [11], as well as a common post processing method (which merges clusters resulting in the biggest cost function gain). This will provide a useful baseline for later comparisons, as well as a principled starting point to improve parallelization efficiency.

In January, we decided to focus on Link Classification in the active setting. Namely, we wanted to build a spanning tree T of the graph and query all its edge signs. In the two clusters case, this allow predicting the sign of $e = (i, j) \in E$ as the product of the signs of edge along the path in T from i to j. Defining the stretch of T as $stretch = \frac{1}{|E|} \sum_{(u,v)\in E} |path^T_{u,v}|$, it turns out that ensuring low error rate amounts to minimizing the stretch, a long open standing problem known as Low Stretch Spanning Tree [26]. Although the theory is not fully ready, experimental results show that our construction is generally competitive with a simple yet efficient baseline and outperforms it for specific graph geometry like grid graphs.

In March, I spent three weeks visiting Claudio Gentile at the Universita' dell'Insubria, Varese, Italy. Professor Gentile was involved in my thesis topic definition and has close links with our research team. We worked on a related problem regarding similarity between the nodes of a graph across different contexts.

2 Roadmap

During the second year, we plan to deepen our understanding of our problems and our methods, by gaining theoretical and experimental insights, which could lead to publications in international workshops or European conferences such as ECML. In the third year, we envisioned more ambitious venues as well as practical applications (e.g. shedding some light on massive real data through our methods). Here are some directions along which we would like to proceed:

KwikCluster proceeds by choosing a distinguished node uniformly at random and putting it in one cluster along with all its positive neighbors, until exhaustion of the graph. A natural extension would be to consider larger neighborhoods, such as nodes at distance at most 2 from the pivot. Although the proof would be more challenging, it could reduce the number of disagreements. Moreover, it would also be interesting to study further parallelization and scaling issues.

In Link Classification, we mentioned that signs can be extended, going from one binary label per edge to a more holistic approach where the similarity between two nodes is measured across different contexts. These contexts are represented by vectors whose dimension matches the dimension of unknown feature vectors associated with each node. The goal is to answer query of the form: how similar are nodes i and j along context $\vec{x}$. We first plan to validate the relevance of this modelling on real problem, then test baseline methods on synthetic and real data before looking for a more effective, online prediction method.

As hinted in the introduction, signed graphs are also amenable to node classification. By leveraging our work on graph sparsifiers, we plan to study node classification by extending the work of Vitale *et al.* [29]. For instance, a fruitful variation of the Low Stretch Spanning Tree problem is to consider a graph where some nodes ($X \subset V$) are distinguished (or revealed) and try to minimize the distances between all pair of points in $X \times \overline{X}$. In the network design community, this is known as the Minimum Cost Routing Tree problem [30, 31]. More generally in Machine Learning, this is an instance of tree based learning [32].

I also plan to work on some follow-ups of my Master Thesis at Aalto about mining urban data.

3 Dissemination policy

As said in the roadmap, my expected production consist solely of publications in Machine Learning and Data Mining conference. In addition, we will consider the opportunity to turn my state of the art about Correlation Clustering into a short survey paper.

4 Professional project

I would rather work in the industry after my PhD. This desire has been reinforced by having taken the self evaluation guide offered by the ABG and the university of Lille. During this year, I had the occasion to talk with members of various companies at ICML and during my summer school, as well as with my adviser, who spent some time in the industry. Later, I plan to apply for the Doctoriales and if it is compatible with Inria policy, do an internship during my third year.

However, it would be wasteful to ignore the academic environment I'm currently working in. Therefore I have followed a training on scientific communication in Nancy and modestly started getting involved in the community by being a volunteer at ICML, helping for the organization of Cap and partially reviewing two papers for NIPS. Finally I will start teaching next year, and I would like to get some formations on this topic as well.

Villeneuve d'Ascq, 09.09.2015

Géraud Le Falher Marc Tommasi

References

[1] M. McPherson *et al.*, "Birds of a feather: homophily in social networks", *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[2] F. Harary, "On the notion of balance of a signed graph", *Michigan Math. J.*, vol. 2, no. 2, pp. 143–146, 1953.

[3] D. Cartwright *et al.*, "Structural balance: a generalization of heider's theory.", *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.

[4] N. Bansal *et al.*, "Correlation clustering", *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pp. 238–247, 2002.

[5] J. Leskovec *et al.*, "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] S. Chawla *et al.*, "Near optimal lp rounding algorithm for correlation clustering on complete and complete k-partite graphs", in *STOC'15*, 2014.

[7] E. D. Demaine *et al.*, "Correlation clustering in general weighted graphs", *Theoretical Computer Science*, vol. 361, no. 2-3, pp. 172–187, 2006.

[8] N. Ailon *et al.*, "Aggregating inconsistent information", *Journal of the ACM*, vol. 55, no. 5, pp. 1–27, 2008.

[9] M. Elsner *et al.*, "Bounding and comparing methods for correlation clustering beyond ilp", pp. 19–27, 2009.

[10] M. Levorato *et al.*, "An ils algorithm to evaluate structural balance in signed social networks", in *Proceedings of the 30th Annual ACM Symposium on Applied Computing - SAC '15*, 2015, pp. 1117–1122.

[11] X. Pan *et al.*, "Scaling up correlation clustering through parallelism and concurrency control", in *NIPS Workshop on Discrete and Combinatorial Problems in Machine Learning (DISCML)*, 2014.

[12] F. Chierichetti *et al.*, "Correlation clustering in mapreduce", in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 641–650.

[13] K. Ahn *et al.*, "Correlation clustering in data streams", in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 2237–2246.

[14] B. Yang *et al.*, "Community mining from signed social networks", *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 10, pp. 1333–1348, 2007.

[15] V. A. Traag *et al.*, "Community detection in networks with positive and negative links", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, 2009.

[16] A. Amelio *et al.*, "Community mining in signed networks", in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013, pp. 95–99.

[17] Y. Chen *et al.*, "Overlapping community detection in networks with positive and negative links", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 3, 2014.

[18] E. W. D. Luca *et al.*, "Spectral analysis of signed graphs for clustering, prediction and visualization", in *Proceedings of the 2010 SIAM International Conference on Data Mining.* 2010, ch. 48, pp. 559–570.

[19] J. Gallier, "Spectral theory of unsigned and signed graphs applications to graph clustering: a survey", 2015.

[20] K.-y. Chiang *et al.*, "Prediction and clustering in signed networks: a local to global perspective", *Journal of Machine Learning Research*, vol. 15, pp. 1177–1213, 2014.

[21] A. Papaoikonomou *et al.*, "Predicting edge signs in social networks using frequent subgraph discovery", *IEEE Internet Computing*, vol. 18, no. 5, pp. 36–43, 2014.

[22] Q. Qian *et al.*, “Efficient distance metric learning by adaptive sampling and mini-batch stochastic gradient descent (sgd)”, *Machine Learning*, vol. 99, no. 3, pp. 353–372, 2014.

[23] J. Ye *et al.*, “Predicting positive and negative links in signed social networks by transfer learning”, in *Proceedings of the 22Nd International Conference on World Wide Web*, 2013, pp. 1477–1488.

[24] N. Cesa-Bianchi *et al.*, “A linear time active learning algorithm for link classification”, in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012, pp. 1–12.

[25] ——, “A correlation clustering approach to link classification in signed networks”, in *Proceedings of the 25th conference on learning theory (COLT 2012).*, 2013, pp. 1–21.

[26] I. Abraham *et al.*, “Using petal-decompositions to build a low stretch spanning tree”, in *Proceedings of the 44th symposium on Theory of Computing - STOC ’12*, 2012, p. 395.

[29] F. Vitale *et al.*, “See the tree through the lines: the shazoo algorithm”, in *Advances in Neural Information Processing Systems 24*, 2011, pp. 1584–1592.

[30] D. S. Johnson *et al.*, “The complexity of the network design problem”, *Networks*, vol. 8, no. 4, pp. 279–285, 1978.

[31] H. S. Connamacher *et al.*, “The complexity of minimizing certain cost metrics for k-source spanning trees”, *Discrete Applied Mathematics*, vol. 131, no. 1, pp. 113–127, 2003.

[32] N. Cesa-Bianchi *et al.*, “Fast and optimal prediction on a labeled tree”, in *COLT*, 2009.