Node Attribute Prediction: An Evaluation of Within- versus Across-Network Tasks

Kristen M. Altenburger[*]
Stanford University
kaltenb@stanford.edu

Johan Ugander
Stanford University
jugander@stanford.edu

Abstract

Node attribute prediction problems arise in a wide range of classification tasks related to graph mining. Examples include detecting bots or spam accounts in online networks or inferring user demographics for targeted marketing. In this work, we evaluate previous research on attribute prediction and discuss the significance of characterizing prediction as *within-network prediction*, where attribute labels are available for nodes in the same network as the unlabeled nodes one is trying to make predictions for, in contrast with *across-network prediction*, where the unlabeled nodes are in a different network than the labeled nodes. We highlight that while much research has focused on within-network prediction, the across-network prediction task can be a more challenging problem than previously appreciated.

Introduction. Predicting node attributes on networks is a problem with a rich history in graph mining. A primary focus for attribute prediction has been on the within-network task [2], which assumes a single network is the population of interest. Within-network attribute prediction has been widely studied, and effective methods include relatively simple approaches based on majority vote algorithms among network neighbors [18], longer-range network aggregation methods based on collective inference [14, 22], LINK methods [17, 29] that employs rows of the adjacency matrix as feature vectors, methods for semi-supervised learning on graphs [16, 30, 31], as well as methods based on embedding-based representations, both supervised (DeepWalk [23], LINE [25], and node2vec [8]) and semi-supervised [28]. These methods for within-network prediction are fundamentally driven by various structural assumptions about how attributes are distributed within a single network of study, assuming that either homophily [20] ("you are similar to the company you keep") or structural equivalence [3, 1] ("you are similar to the company you're kept in") governs some extent of how individuals relate to each other in the network. As such, these methods all base their predictions on similarities (of some variety) along edges or paths, and are explicitly not identifying innate structural features of nodes with given attributes.

In across-network prediction, one or more networks are accompanied by a complete set of node attribute labels, but the goal is to predict attribute labels for nodes on another network (or set of networks). Therefore, across-network prediction eliminates the possibility of relating unlabeled nodes to labeled nodes along paths of any length. The only methods that are admissible for across-network prediction are ones that learn relationships between attribute labels and innate structural features of nodes: how many friends they have, their clustering coefficient, or richer representations. Informally, it requires discovering something about the way the nodes in a labeled training network G_1 are positioned that also holds true in an unlabeled testing network G_2. We observe that previous benchmarks for across-network tasks have been evaluated on easier within-network problems, with limited evaluations of genuine across-network problems.

In order to appreciate the within-network versus across-network attribute prediction distinction, we begin by introducing related concepts of label-dependent and label-independent feature representations [6]. A *label-dependent* node feature is a feature that depends on the labels or attributes of nodes

[*] KMA was supported in part by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship program.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

in the graph, whereas a *label-independent* node feature is a feature that does not. Clear examples of label-dependent features of a node v_i include the number of female friends of v_i or the distance from v_i to nearest male node. Clear examples of label-independent features include the degree of v_i or the number of triangles containing v_i. But not all features are so easily classified as label-dependent or independent. For example, consider LINK features [29], where rows of the graph adjacency matrix are employed as a large sparse feature vector. LINK features coupled with regularized logistic regression has been found to be highly effective when deployed for various attribute prediction tasks. These LINK features are label-independent in the sense that they do not depend on any node attributes, but LINK features do, however, depend on the *identity labels* $1, \ldots, n$ of the nodes.

In this work we consider a feature to be label-independent if it is invariant to arbitrary re-labelings of the node set, and label-dependent otherwise. It is clear that label-dependent methods are only useful for within-network tasks: features such as "number of female friends" or "is friends with node 1" clearly can't be translated from a source graph G_1 to a target graph G_2. But notice that label-dependent methods are admissible for across-layer tasks (in multi-layer networks [15]) because the node set is the same. We highlight that across-network studies in social network settings have in fact been evaluated on across-layer tasks, and find that across-layer tasks overstate the predictive performance of label-independent features for across-network tasks.

Datasets. We illustrate the distinctions between within-network and across-network tasks in the context of two main datasets: the Facebook100 (FB100) dataset [26, 27] for gender classification and the Reality Mining dataset [5] for student status classification. The FB100 dataset consists of the online friendship networks from the first 100 colleges that accessed the Facebook platform in 2005, as released by Facebook, and includes gender, class year, and other attributes. The FB100 dataset lends itself to across-network tasks as we have networks across different college settings, and also lends itself to within-network tasks where we can treat a specific school as the population of interest.

The Reality Mining dataset consists of data from a study at MIT in 2004-2005 that tracked the cell phone usage of 94 subjects comprised of students and faculty. The subjects agreed to have their interactions and proximity to one another recorded over the course of the study period. One can conceive of different network attribute prediction tasks such as inferring whether a participant is a graduate student in the Media Lab or is a business school student [10]. It is possible to view this dataset through either (a) the lens of across-network prediction by treating interactions occurring in different months as comprising different networks or (b) the lens of within-network prediction by treating the full network of 94 subjects and their interactions over time as the network of interest. We consider the proximity network from the Reality Mining dataset with the goal of inferring subject-level attributes to be a within-network prediction task as there is only one node set, subjects who participated in the study. The dataset has historically been sliced across time for the purpose of creating a plausible across-network task, but this framing leads to a very different across-layer task compared to settings where the target node population is genuinely from a different network.

Within-network prediction is easy. For the cross-validation set-up in a within-network task, we vary the percent of nodes initially labeled and evaluate across random sets of nodes that are selected as training data.[2] The set of labeled nodes are selected i.i.d., implementing a missing completely at random (MCAR) [9] missingness mechanism. Within this cross-validation set-up, one could also consider varying the type of missingness mechanism generating which node attributes are observed and which are private, but we do not explore such mechanisms here.

LINK-based models can leverage friend-of-friend information to achieve high predictive performance in the presence of monophily [1], an analog of homophily that corresponds to structural equivalence. We observe that LINK achieves high performance, as illustrated in Figure 1, both for predicting student status (MIT Sloan business school students vs. not) in the Reality Mining dataset (aggregating interactions across the full observation period), and gender in Amherst College (a representative sample network from the FB100 dataset). However, as we've discussed, LINK is a label-dependent feature representation that uses identity labels, making it network-specific. As such it cannot be applied in across-network settings. We therefore investigate how well we can do on this within-network task using label-independent feature representations, specifically ReFeX [11].

[2] Neville et al. [21] recommend an alternative sampling approach aimed at maintaining correct Type I error rates when comparing model performance using paired t-tests, which we do not employ here.

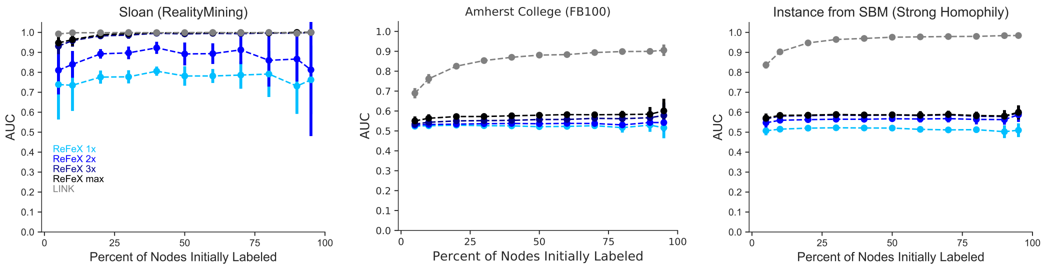

Figure 1: Within-Network Prediction: We contrast the high predictive performance of inferring whether a student is a business school student (left) versus the limited predicted performance of inferring gender in the FB100 dataset on Amherst College (center) using label-independent features (ReFeX). Meanwhile, on a network instance from a highly homophilous stochastic block model (SBM), we observe LINK has high predictive performance while recursive features are limited due to the lack of a structural signal (right).

Using ReFeX features we train a LogForest model, an ensemble of logistic regression models, as used in the original ReFeX work [7]. We compute ReFeX features several different ways. Specifically, we vary the recursion depth of the feature generating process, extracting representations based on 1, 2, 3, and a maximum (up to 100) number of recursive iterations. The "max" recursion representations are based on only a handful of recursions, far less than 100, because of the feature pruning mechanism built into the (standard) ReFeX implementation we employed[3]. For predicting student status, we observe high performance of the ReFeX features, suggesting that business school student status is structurally distinct in this network. However, for gender prediction on Amherst College we observe extremely limited performance based on ReFeX features.

Across-network prediction is deceptively hard. We illustrate that across-network prediction can be a more difficult problem by comparing the across-time prediction for business school student status (Reality Mining) versus the across-school prediction for gender (Amherst from FB100). The across-network predictions of student status are based on viewing the Reality Mining data as a time-sliced network with each month corresponding to a different slice. From month to month we are still making predictions on the same node set, and arguably label-dependent features such as LINK are fully admissible, but we restrict ourselves to label-independent features to understand their capability in this setting. The across-network gender prediction task in the FB100 networks takes place across networks with disjoint node sets.

Cross-validation for across-network tasks have different considerations than for within-network tasks. First, the type of node attribute missingness mechanism is not a concern at the node-level since we observe all labels for a graph. However, a different missingness concern is this setting is which graphs do we observe? For generating ReFeX features, we note that the node representations are trained during separate ReFeX iterations, which can result in features being (a) binned differently and (b) have different feature sets selected by the pruning mechanism. To address the first challenge, we normalize all features to be in $[0, 1]$. To address the second challenge, we employ a "double pass" routine to find all the features selected by the pruning mechanism across all networks in the collection in a first pass, and then repeat the feature extraction while manually requiring ReFeX to return the union of all these features as the representation.

We again consider ReFeX feature sets generated by 1, 2, 3, and a maximum number of recursive iterations, allowing us to identify at which recursion step performance gains occur. For predicting student status in the Reality Mining dataset, we follow earlier work by using consecutive months in a paired train/test set-up [10]. For predicting gender in the FB100 dataset, we use Amherst College as the train school and compare performance when using different schools for testing. Different training schools give comparable results. As illustrated in Figure 2, we observe slight performance improvement at higher recursions when predicting business student status, though the main performance improvement (over the baseline) comes from the ReFeX base representation ("ReFeX 1x"), before any aggregation functions have been applied. For gender prediction we observe

[3] https://github.com/LLNL/refex-rolx

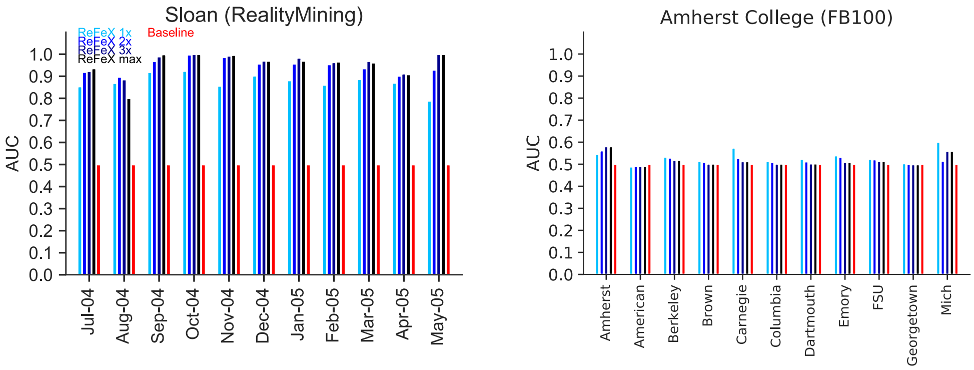

Figure 2: Across-Network Prediction: We demonstrate on both predicting business student status across time in the Reality Mining data (top) and gender across schools in the FB100 dataset (bottom) see only modest predictive performance gain at deeper levels of recursion.

stable, low predictive performance across all recursive depths. Put simply, there is little gain in increasing the number of iterations when creating ReFeX features.

These simple examples demonstrate instances when within-network prediction can be high (e.g., LINK in Amherst College) but where across-network prediction performance is much more limited. That is, high predictive performance for within-network tasks using label-dependent features do not imply high performance for the across-network task using label-independent features. We notice in the earlier Figure 1 that the ReFeX feature set had very limited performance for within-network prediction on Amherst. Therefore, it's not surprising that there's limited performance from label-independent features in the across-network setting as well.

A challenge with interpreting the limited across-network performance on Amherst College is revealed when considering attribute prediction on graphs generated from a stochastic block model (SBM) [13]. Consider a strongly homophilous SBM with two blocks of equal size (n=2000 nodes in total, average degree of 84, and with homophily index [4] of approximately 0.60). We evaluate the within-network performance of LINK and ReFeX features on such a network in Figure 1 (right) and observe a story similar to the within-network prediction on Amherst College: LINK performs very well, able to leverage the similarity among *individuals* that serve as useful features. Meanwhile ReFeX is not able to leverage any structural signal because, from the perspective of ReFeX, nodes in the two blocks are identical. We see that ReFeX and related label-independent features are blind to simple, strong, but symmetric structures that do not translate easily to label-independent node features.

For the research community to make progress on addressing across-network problems for node attribute prediction, especially in settings when node sets may be completely disjoint, we must first agree on what task a proposed node attribute method is being tested on, benchmark datasets for evaluation, and how to measure performance. The challenges identified in this work suggest productive avenues for future research on node attribute prediction. We believe it is a synergistic opportunity to critically examine other social science disciplines that have been wrestling with similar across-network type problems. For example, the *ideal points* literature in American politics pursues a structural approach to rank judges across time based on their voting behavior [24, 19]. The ideal points estimation literature has already thought critically about issues such as temporal extrapolation [12], and there is likely potential to learn from and adapt some of these approaches.

Despite the impressive performance of LINK for the within-network problem, it is important to note that relational inference requires training a model specific to the network in question. The high performance of LINK is potentially suggestive of overlooked novel label-independent features. As an open question, our analysis provides no theoretical basis for limiting the predictive performance of models based on ReFeX features (or any other label-independent features) relative to performance with LINK, and it is possible that the "right" label-independent features enable across-network predictions with performance on par with the within-network predictions based on LINK features. How should we reconsider how we think about across-network tasks? What does the transferability of a model between pairs of networks say about those networks in a larger population? A first step forward is to define a clear consistent set of across-network tasks, with many challenges to follow.

References

[1] Kristen M Altenburger and Johan Ugander. Monophily in social networks introduces similarity among friends-of-friends. *Nature Human Behaviour*, 2(4):284, 2018.

[2] Smriti Bhagat, Graham Cormode, and S Muthukrishnan. Node classification in social networks. In *Social Network Data Analytics*, pages 115–148. Springer, 2011.

[3] Ronald S Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):1287–1335, 1987.

[4] James Coleman. Relational analysis: the study of social organizations with survey methods. *Human Organization*, 17(4):28–36, 1958.

[5] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[6] Brian Gallagher and Tina Eliassi-Rad. Leveraging label-independent features for classification in sparsely labeled networks: An empirical study. In *Advances in Social Network Mining and Analysis*, pages 1–19. Springer, 2010.

[7] Brian Gallagher, Hanghang Tong, Tina Eliassi-Rad, and Christos Faloutsos. Using ghost edges for classification in sparsely labeled networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–264. ACM, 2008.

[8] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[9] Daniel F Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.

[10] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. RolX: structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1231–1239. ACM, 2012.

[11] Keith Henderson, Brian Gallagher, Lei Li, Leman Akoglu, Tina Eliassi-Rad, Hanghang Tong, and Christos Faloutsos. It's who you know: graph mining using recursive structural features. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 663–671. ACM, 2011.

[12] Daniel E Ho and Kevin M Quinn. How not to lie with judicial votes: Misconceptions, measurement, and models. *California Law Review*, 98(3):813–876, 2010.

[13] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[14] David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. In *Proceedings of the Tenth ACM SIGKDD International Cconference on Knowledge Discovery and Data Mining*, pages 593–598. ACM, 2004.

[15] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[16] Danai Koutra, Tai-You Ke, U Kang, Duen Horng Polo Chau, Hsing-Kuo Kenneth Pao, and Christos Faloutsos. Unifying guilt-by-association approaches: Theorems and fast algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 245–260. Springer, 2011.

[17] Qing Lu and Lise Getoor. Link-based classification. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 496–503, 2003.

[18] Sofus A Macskassy and Foster Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, 2007.

[19] Andrew D Martin and Kevin M Quinn. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002.

[20] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

[21] Jennifer Neville, Brian Gallagher, Tina Eliassi-Rad, and Tao Wang. Correcting evaluation bias of relational classifiers with network cross validation. *Knowledge and Information Systems*, 30(1):31–55, 2012.

[22] Jennifer Neville and David Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20, 2000.

[23] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710. ACM, 2014.

[24] Keith T Poole and Howard Rosenthal. *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand, 2000.

[25] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[26] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.

[27] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[28] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, pages 40–48, 2016.

[29] Elena Zheleva and Lise Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, pages 531–540, 2009.

[30] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.

[31] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.