Decoupled Smoothing in Probabilistic Soft Logic

Yatong Chen*
ychen592@ucsc.edu
University of California, Santa Cruz
Santa Cruz, USA

Bryan Tor*
btor@ucsc.edu
University of California, Santa Cruz
Santa Cruz, USA

Eriq Augustine
eaugusti@ucsc.edu
University of California, Santa Cruz
Santa Cruz, USA

Lise Getoor
getoor@ucsc.edu
University of California, Santa Cruz
Santa Cruz, USA

ABSTRACT

Node classification in networks is a common graph mining task. In this paper, we examine how separating *identity* (a node's attribute) and *preference* (the kind of identities to which a node prefers to link) is useful for node classification in social networks. Building upon recent work by Chin et al. (2019), where the separation of identity and preference is accomplished through a technique called "decoupled smoothing", we show how models that characterize both identity and preference are able to capture the underlying structure in a network, leading to improved performance in node classification tasks. Specifically, we use probabilistic soft logic (PSL) [2], a flexible and declarative statistical reasoning framework, to model identity and preference. We compare our approach with the original decoupled smoothing method and other node classification methods implemented in PSL, and show that our approach outperforms the state-of-the-art decoupled smoothing method as well as the other node classification methods across several evaluation metrics on a real-world Facebook dataset [24].

ACM Reference Format:
Yatong Chen, Bryan Tor, Eriq Augustine, and Lise Getoor. 2020. Decoupled Smoothing in Probabilistic Soft Logic. In *MLG '20: International Workshop on Mining and Learning with Graphs, Aug 24, 2020 - San Diego, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Classifying, or labeling, nodes in networks is a common graph mining task for which a wide variety of methods have been proposed [7, 9, 13, 15, 16, 23, 28, 29]. Most methods infer information about a node's label based on its attributes, relational structure, and neighbors' labels. Many methods also propagate node labels along edges in order to jointly infer unobserved labels. Within social networks, the phenomenon of *homophily*, where neighboring nodes tend to have the same label [19], is commonly exploited. Another phenomenon, known as *monophily*, where nodes share similarity with their neighbors' neighbors, has also been shown to be useful in identifying unknown labels [1].

In this paper, we examine how the notions of *identity* and *preference* are useful for node classification in social networks. Identity refers to a specific attribute of an individual, such as their gender, political affiliation, religious belief, and so on; preference refers to the tendency of an individual to share a connection with others having a particular identity. The idea of decoupling *identity* and *preference* was first introduced by Chin et al. (2019). In their work, they make specific assumptions on the correlations between an individual's identity and preference: they assume that the identity of a given node, an *ego* node, is approximated by a weighted average of its neighbors' preferences, and the preference of the ego node can be approximated by a weighted average of its neighbors' identities. Following these assumptions, all preferences can be eliminated from the system, so the unlabeled identities can be inferred without explicitly modeling the preference of any node. We refer to this approach as "original decoupled smoothing" (DS ORIG).

We build upon the original decoupled smoothing work and show how to use probabilistic soft logic (PSL) [2], a statistical relational learning framework for relational domains, to model preferences and identities. We show that PSL can solve node classification problems efficiently using the concept of decoupled smoothing. In addition, the rich modeling capabilities of PSL allow us to incorporate prior information and domain knowledge into our model.

We perform an empirical study using different approaches on a real Facebook dataset [24]. Specifically, we compare the PSL implementation of Decoupled Smoothing (DS PSL) with the original Decoupled Smoothing method (DS ORIG) and other existing classification methods based on homophily (1-HOP PSL) or monophily (2-HOP PSL), by applying them to a gender labeling task. Our results first show that DS PSL outperforms 1-HOP PSL, 2-HOP PSL, and DS ORIG in terms of both categorical accuracy and AUROC, especially when less than 50% percent of the node labels are observed. This indicates that DS PSL is able to better capture the underlying network structure by modeling identity and preference explicitly, especially when the label information is sparse. In addition, we find that while DS ORIG fails to outperform 2-HOP PSL, DS PSL outperforms 2-HOP PSL, showing the effectiveness of decoupled smoothing in PSL as a fundamental modeling tool. We also explore a variation of DS PSL by adding an additional rule that incorporates local homophily properties of preference among individuals that are tightly connected with each other. This achieves similar performance compared to DS

*Both authors contributed equally to this research.

MLG '20, Aug 24, 2020, San Diego, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

PSL, but provides a way to exploit additional relational structures in the graph beyond friendship links.

2 PROBLEM STATEMENT

In this work, we focus on label prediction in social networks. Given the structure of a social network graph and the labels for a subset of nodes, the task is to infer the labels for the unlabeled nodes. Specifically, we assume that we are given a social network of individuals, represented as nodes, with social ties, such as friendship, represented as undirected, unweighted edges. The *neighbors* of node i refers to the set of individuals who are immediate friends with i. Each person i is associated with an identity and a preference. Identity is the label that we are trying to predict, and corresponds to a specific attribute of an individual, such as gender, political affiliation, or religious affiliation. In contrast, the preference of an individual is their tendency to have social ties with individuals of a certain identity. Preference is completely unobserved and treated as a latent variable

The separation of identity and preference is more evident when the preference is not *homophily-driven*. For example, users of similar political affiliation tend to prefer neighbors of that same affiliation, which results in a user's identity and preference being the same. Figure 1a shows that the center node's political affiliation is Party A, and most of its friends also have the same political affiliation of Party A. In this case, the separation of identity and preference is *less* obvious, since preference of political affiliation is homophily driven. However, users of a particular gender may not always prefer neighbors of that same gender, which results in users having a preference that is not the same as their identity. As is shown in Figure 1b, the center node's gender is male. However, it has a preference of making friends with people whose identities are female. In this case, the separation of identity and preference is *more* apparent, since gender preference is not always homophily driven.

In this work, we focus on the specific task of gender prediction. In this setting, a person's identity is their gender and their preference is a latent variable indicating their tendency to make friends with a particular gender. Because of limitations in the dataset, we treat gender as a binary label: *Female* or *Male*.

3 BACKGROUND

In this section, we provide a brief review of the properties of large-scale social networks, decoupled smoothing on graphs, and Probabilistic Soft Logic (PSL).

3.1 Properties of Large-Scale Social Networks

The recent emergence and popularization of Online Social Networks (OSNs) has made available a large amount of data on social organization, interaction, and human behavior, providing many research opportunities for data mining in large-scale networks. Node classification is a common graph mining task. There are several special characteristics that are commonly observed in large-scale social network graphs that can help improve the accuracy of predictions. First is the well-known phenomena of *homophily* [14], in which individuals tend to be connected with people who are similar to themselves. This phenomena is sometimes referred to as "birds of a feather flocking together", and is often observed in social networks [19]. For neighbors within a network, homophily finds similarity between their attributes [6]. Another phenomenon that often exists in large-scale social networks is *monophily* [1]. Monophily is the phenomena where attributes of an individual's friends are likely to be similar to the attributes of the individual's other friends. As pointed out by Altenburger and Ugander (2018), in cases where homophily is weak or nonexistent, monophily has been shown to still hold.

In addition to these statistical characteristics of labels, social networks also exhibit various purely topological properties. In most real-world OSNs, people are likely to form highly clustered communities, which is one of the distinguishing features of social networks [26]. The degree distribution is highly skewed [22]. Moreover, many graphs have high clustering coefficients [27], which is indicative of underlying community structure.

3.2 Decoupled Smoothing on Graphs

Decoupled smoothing was first introduced by Chin et al. (2019). The key principle of decoupled smoothing is the separation of a person's identity from their preference. Suppose we have an undirected, unweighted social network graph and an associated matrix W, where each element W_{ij} represents the influence of individual i on individual j. W_{ij} is non-zero if i and j are friends. Let the row sums be denoted by $z_i = \sum_j W_{ij}$, and the column sums be denoted by $z'_j = \sum_i W_{ij}$. Decoupled smoothing relates an individual i's identity θ_i and preference ϕ_i via W as follows:

$$\theta_i \approx \frac{1}{z_i} \sum_{j=1}^{n} W_{ij}\phi_j$$

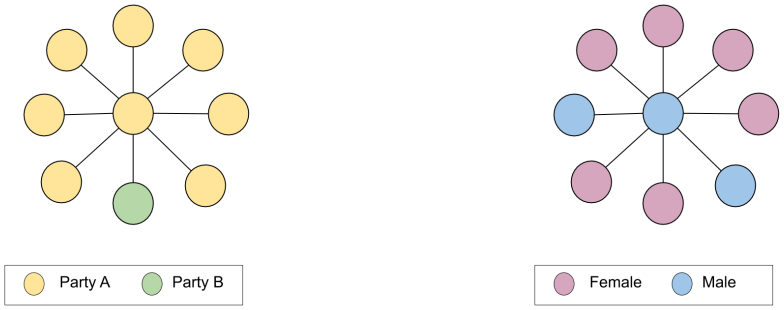
$$\phi_j \approx \frac{1}{z'_j} \sum_{i=1}^{n} W_{ij}\theta_i$$

Intuitively, this means that individual i's identity θ_i is a weighted average of its friends' preferences, and its preference ϕ_i is a weighted average of its friends' identities. Chin et al. (2019) show that this model is equivalent to marginally specifying the joint Gaussian distribution for θ and ϕ. Since the goal is to obtain predictions for unobserved identities, they view the preference variable as a nuisance parameters and marginalize them out. In this way, preferences can be eliminated from the system while its information can still be encoded in the remaining linear system. The authors then propose various ways to estimate the weight matrix W under certain assumptions. The result is that node identities can be inferred without explicitly modeling the preference of any node.

Notions of identity and preference are not limited to social networks and the idea can be applied to any attributed graph.

3.3 Probabilistic Soft Logic (PSL)

Probabilistic Soft Logic (PSL) is a statistical reasoning framework for collective, probabilistic reasoning in relational domains. A PSL model is defined through a set of weighted first-order logical rules, which can be used to specify features of graphical models over ground atoms with a continuous relaxation of Boolean logic. We point the reader to Bach et al. (2017) for a more detailed discussion



a. An example of political affiliation identity and preference. The center node's political affiliation is Party A, and most of its friends share the same political affiliation. In this case, identity and preference share a large overlap, since the preference of political affiliation is homophily driven.

b. An example of gender identity and preference. The center node's gender is male, but most of its friends' are female. In this case, identity and preference are almost disjoint, since the preference of gender is not always homophily driven.

Figure 1: A demonstration on decoupling identity and preference for political affiliation identity and gender identity.

of PSL. Here, we illustrate the key ideas by providing an example:

$$w : \textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$$

In this PSL rule, $w \in R^+$ is a learnable weight, indicating the importance of satisfying this rule. Edge and Gender are two predicates, where A/B and G are placeholders for a person and a gender, respectively. In our setting, the Edge predicate takes two nodes as its arguments, and represents the friendship link between two people. When the value of this predicate is 1, we determine that the two nodes are friends in the social network, and 0 otherwise. The value of a predicate can also be a numeric value in $[0, 1]$. When a rule is instantiated with data, e.g.,

$$w : \textsc{Edge}(\text{"Alice"}, \text{"Bob"}) \wedge \textsc{Gender}(\text{"Alice"}, \text{"Female"}) \rightarrow \textsc{Gender}(\text{"Bob"}, \text{"Female"})$$

it is referred to as ground rule and each atom in a ground rule, such as Gender("Alice", "Female"), is referred to as a ground atom. Each ground atom is represented as a continuous variable in the range of $[0, 1]$ and each ground rule represents a clique in hinge-loss Markov random field (HL-MRF). Given the observed variables, X, and unobserved variables, Y, the probability density of a HL-MRF is:

$$P(Y|X) \propto \exp\Big(-\sum_{i=1}^{m} w_i \phi_i(Y, X)\Big),$$

$$\text{where } \phi_i = \max\{0, \ell_i(Y, X)\}^{d_i},\ d_i \in \{1, 2\}$$

where m is the total number of cliques, ϕ_i is a potential function associated with a clique generated by a ground rule, ℓ_i is a linear function, d_i is the choice between linear and squared hinge loss (we use squared in this paper), and w_i is the weight associated with the rule. The task of inference can then be written as:

$$\arg\max_Y P(Y|X) = \arg\min_Y \sum_{i=1}^{m} w_i \phi_i(Y, X)$$

The above expression can be solved using the Alternating Direction Method of Multipliers (ADMM) [5].

4 METHODOLOGY

In this section, we describe different gender prediction models in PSL. For all of our models, Edge(A, B) is fully observed. The Gender predicate takes two arguments: a node A and a gender G. For all of our models, Gender is partially observed. Gender(A, "Female") represents the probability that person A is female, and Gender(A, "Male") is the probability that person A is male. For those whose genders are observed, the value of the atom that corresponds to their true gender is 1, and the value of the atom for the other gender is 0. For those whose genders are unobserved, we will infer their two gender predicates Gender(A, "Male") and Gender(A, "Female") jointly, and then assign the predicted gender based on the majority value.

For all models, we apply a functional constraint to the Gender predicate. This constraint ensures that sum of the female and male atoms is always 1.

$$\textsc{Gender}(\texttt{A}, +\texttt{G}) = 1$$

We list all the rules associated with each model in Table 1. Next, we will explain each method and their corresponding PSL rules in details.

One-hop Method. The one-hop method (1-hop psl) relies solely on homophily. The PSL implementation of one-hop is accomplished in one rule: if two nodes A and B share an edge, and A has the gender attribute G, we conclude that node B is likely to have the gender attribute G as well:

$$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$$

Two-hop Method. The two-hop method (2-hop psl) is based on monophily. This method uses the relationship between a node and its two-hop neighbors. Like one-hop, the PSL implementation of

Model	PSL Rules
1-hop psl	$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$ $\textsc{Gender}(\texttt{A}, +\texttt{G}) = 1$
2-hop psl	$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Edge}(\texttt{B}, \texttt{C}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{C}, \texttt{G})$ $\textsc{Gender}(\texttt{A}, +\texttt{G}) = 1$
ds psl	$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Preference}(\texttt{B}, \texttt{G})$ $\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Preference}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$ $\textsc{Preference}(\texttt{A}, +\texttt{G}) = 1$ $\textsc{Gender}(\texttt{A}, +\texttt{G}) = 1$
ds-pc psl	$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Preference}(\texttt{B}, \texttt{G})$ $\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Preference}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$ $\textsc{CloseFriend}(\texttt{A}, \texttt{B}) \wedge \textsc{Preference}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Preference}(\texttt{B}, \texttt{G})$ $\textsc{Preference}(\texttt{A}, +\texttt{G}) = 1$ $\textsc{Gender}(\texttt{A}, +\texttt{G}) = 1$

Table 1: PSL rules for different models.

2-hop psl only requires one rule: if three nodes A, B, and C form a relationship chain such that A is friends with B, and B is friends with C, then we conclude that nodes A and C are likely to have the same gender attribute:

$$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Edge}(\texttt{B}, \texttt{C}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{C}, \texttt{G})$$

Decoupled Smoothing. The decoupled smoothing (ds psl) model allows an individual's gender preference to differ from their own gender identity. In order to achieve this, we add a Preference predicate, representing each person's propensity to befriend people of a particular gender. $\textsc{Preference}(\texttt{A}, \texttt{G})$ can any take any value within the range of $[0, 1]$: a value of 1 implies A strongly prefers friends of gender G, a value of 0 implies A strongly prefers friends not of gender G, and any value between falls on that spectrum. Unlike gender, there is no explicit information available for preference. In our approach, we learn a person's preference by jointly reasoning about both their identity (gender) and preference: if two nodes A and B share an edge, and A has gender attribute G, then we conclude that B likely has a preference for gender attribute G. Furthermore, if two nodes A and B share an edge, and A has a preference for gender attribute G, then we conclude that B likely has a gender attribute G. The corresponding PSL rules are:

$$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Gender}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Preference}(\texttt{B}, \texttt{G})$$
$$\textsc{Edge}(\texttt{A}, \texttt{B}) \wedge \textsc{Preference}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Gender}(\texttt{B}, \texttt{G})$$

We apply a functional constraint to the Preference predicate:

$$\textsc{Preference}(\texttt{A}, +\texttt{G}) = 1$$

Decoupled Smoothing with Preference Concentration. Next, we introduce a model that captures both preferences and community structure. We make an additional assumption that a pair of friends who share a large number of common friends are more likely to also share similar preferences. We refer to this method as "decoupled smoothing with preference concentration" (ds-pc psl). To measure how closely two friends are related, we create an observed CloseFriend predicate, which takes two individuals as its arguments, and represents the "closeness" between them based on the number of their common friends. We then measure "closeness" either by normalizing the number of common friends (divide it by the largest number of common friends in the graph) shared by individuals, or with a threshold. Figure 2 show how different thresholds can lead to a significant difference in the number of common friends a pair of individuals share. However, empirical results show that different threshold choices do not lead to a significant change in performance. As a result, we will use 200 as a representative threshold of ds-pc psl since it takes the shortest amount of time to run. We will address it as ds-pc psl (200). We also consider the normalized version of CloseFriend, denoted as ds-pc psl (normalized).

In addition to the rules established for decoupled smoothing, we add a rule representing the idea that two individuals with a higher value of CloseFriend are more likely to have a similar Preference. This additional rule allows the model to focus on pair of individuals who belong to the same clustered local communities, where *homophily* on preference might be stronger.

The corresponding additional PSL rules are:

$$\textsc{CloseFriend}(\texttt{A}, \texttt{B}) \wedge \textsc{Preference}(\texttt{A}, \texttt{G}) \rightarrow \textsc{Preference}(\texttt{B}, \texttt{G})$$

5 EMPIRICAL EVALUATION

In this section, we evaluate the performance of 1-hop psl, 2-hop psl, ds orig, ds psl, and ds-pc psl on a gender classification task using a real-world Facebook dataset [24].

5.1 Dataset and Evaluation Metric

Our datasets consists of a network of Facebook users who were undergraduates attending Amherst College in 2005 [24]. We use the largest connected component from the network, which contains 2032 nodes and 78733 edges, with 1017 female users and 1015 male users. We only consider nodes that have a self-reported gender. Following the evaluation of Chin et al. (2019), we uniformly sample

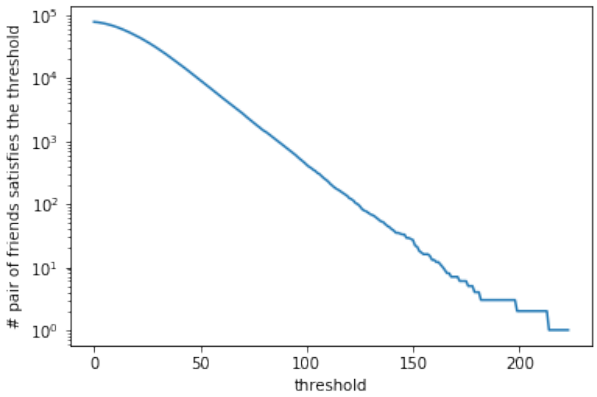


Figure 2: The number of pairs of friends whose number of common friends is greater than a threshold. Note that the y-axis uses a log scale.

a set of the nodes to be labeled initially, and evaluate on the remaining unlabeled nodes. The percentage of nodes that are initially labeled ranges from 1% to 90%. The sampling process is repeated ten times to create ten independently sampled splits for each labeling percentage.

To evaluate the performance of our methods on gender (identity) prediction, we measure classification performance using both AUROC and categorical accuracy. We also reproduce the experiment conducted by Chin et al. (2019) to obtain the original decoupled smoothing result for comparison.

5.2 Results

Figure 3 shows the AUROC and categorical accuracy for each method across different percentages of initially labeled nodes. We show the average values across ten trials. Table 2 shows the AUROC macro-average across all labeling percentages as well as the AUROC for 20%, 50%, and 90% labeled nodes with standard deviations included.

First, we will discuss the difference in performance between 1-hop psl and 2-hop psl. In terms of AUROC, 2-hop psl outperforms 1-hop psl regardless of the percentage of initially label nodes. In terms of categorical accuracy, 1-hop psl performs better when less than 50% of the nodes are initially labeled, but then 1-hop psl performs worse when more than 50% of the nodes are initially labeled. Overall, 1-hop psl has a lower standard deviation compared to 2-hop psl for both AUROC and categorical accuracy. This indicates that when observed labels are sparse, homophily is more useful than monophily. This is possibly because when the observed labels are sparse, less two-hop neighbors will be observed, thus it is difficult to make use of the monophily.

Next, we observe that ds psl outperforms ds orig across all metrics and labeling percentages, with the gap in performance growing as fewer nodes are initially labeled. This indicates that explicitly modeling preference and performing joint inference allows ds psl to better capture the underlying interaction between identity and preferences among individuals, especially when the observed labels are sparse. ds orig, however, heavily relies on specific assumptions made on the interaction between identity and preferences for each individual and their neighbors.

Third, ds psl outperforms 2-hop psl, while ds orig preforms similarly to 2-hop psl. This supports our assertion that decoupled smoothing is a more expressive model than 2-hop psl, and unlike 2-hop psl , both ds psl and ds orig do not directly rely on the monophily phenomena.

We also observe that ds-pc psl performs similarly to ds psl despite incorporating additional information from the CloseFriend predicate. This is likely because of weak shared community preferences in our data. We believe that ds-pc psl may have better performance if evaluated on a network where there is higher shared community preferences, or in a setting where the preference is more homophily driven.

6 RELATED WORK

There has been extensive research on node classification for large scale social networks. Existing node classification techniques mainly fall into the following two categories [3]: methods based on iterative application of traditional classifiers using graph information as features, and methods which propagate the existing labels using random walks. The algorithms which belong to the first category iteratively predict labels for the unlabeled nodes in the graph using labels predicted in the previous iteration. For example, as discussed in the methodology section, Macskassy and Provost (2003) introduced the weighted-vote relational neighbor classifier, which is based on a direct application of homophily. Lu and Getoor (2003) introduced link-based classification, which proved to be a monophily based method by [1]. The second type of method relies on performing some forms of label propagation on the social network graphs based on random walk methods. Examples include the work of Zhu et al. (2003) on semi-supervised learning using Gaussian Markov random fields, and Zhou et al.'s (2004) related method for random walk smoothing. The original decoupled smoothing work of Chin et al. (2019) belongs to this category as well. In addition, there is a fair amount of recent work on the use of higher-order network structures that aid graph semi-supervised learning [9] and deep generative models [8, 12, 13].

In addition to PSL, other statistical relational learning (SRL) methods have been used for network classification tasks because of their ability to exploit relationships between labels of related nodes. Sen et al. (2008) provides a survey of collective classification in network data. Neville and Jensen (2000) present an iterative classification procedure that exploits relational data, which is the first method in the SRL community on collective classification. McDowell and Aha (2012) examine how to improve the semi-supervised learning of collective classification models when given only a sparsely labeled graph. Gallagher et al. (2008) propose a "ghost edge" method to work for scenarios where homophily may not necessarily hold for a network, which is also one of the key motivations for the development of decoupled smoothing. Ghamrawi and McCallum (2005) explores multi-label conditional random field classification models that directly parameterize label co-occurrences in multi-label classification settings. Bilgic et al. (2007) introduces a way to combine collective classification and link prediction. Moore and Neville (2017)

Model Name	Average	20%	50%	90%
1-HOP PSL	0.640±0.021	0.637±0.019	0.675±0.011	0.659±0.047
2-HOP PSL	0.689±0.034	0.670±0.038	0.730±0.034	0.738±0.042
DS ORIG	0.688±0.020	0.663±0.036	0.724±0.029	0.740±0.041
DS PSL	**0.703±0.030**	**0.698±0.031**	**0.746±0.035**	**0.754±0.034**
DS-PC PSL (normalized)	**0.701±0.030**	**0.703±0.032**	**0.752±0.030**	**0.755±0.033**
DS-PC PSL (200)	**0.699±0.030**	0.683±0.029	**0.742±0.032**	**0.755±0.033**

Table 2: Experimental results showing the AUROC averaged over all labeling percentages, and with 20%, 50%, and 90% of nodes labeled. The significantly best results, at p = 0.05, are shown in bold. All PSL-based decoupled smoothing methods beat out all the non-PSL methods.

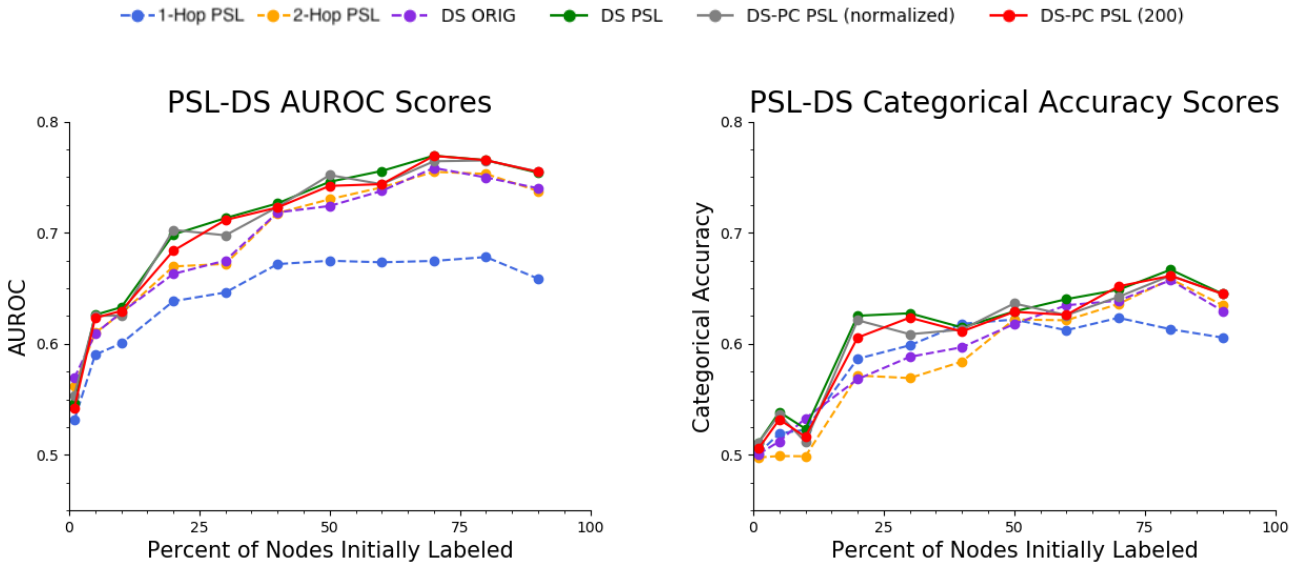


Figure 3: AUROC and categorical accuracy on all methods averaged over 10 samples with the percentage of initially labeled nodes ranging from 5% to 90%.

exploits recent development in recurrent neural networks (RNN) for collective inference classification in network datasets.

7 CONCLUSION AND FUTURE WORK

In this work, we study the modeling of *identity* and *preference* for node classification in social networks using PSL. Building on work by Chin et al. (2019), we propose DS PSL, an implementation of decoupled smoothing in PSL. Unlike previous work, which makes strong assumptions about the correlation of identity of an individual with the preference of their neighbors, DS PSL is able to avoid this by modeling both the identity and preference of an individual. We also implement other node classification methods which explicitly model the correlation between the identities of individuals and their neighbors (1-HOP PSL, 2-HOP PSL) in PSL. We evaluate these methods on a real-life Facebook dataset for a gender classification task. Our results demonstrate that DS PSL is able to achieve better classification performance than state of art (DS PSL) and other models in terms of AUROC and categorical accuracy, especially when the initially observed labels are sparse. This shows that decoupled smoothing in PSL is better at capturing the underlying network structures without making additional assumptions on the specific interactions between preference and identity between individuals and their neighbors.

Because of the flexible nature of decoupled smoothing in PSL, there are many opportunities for further improvements. First, we can easily incorporate external information on either identity or preference in the form of additional PSL rules. In the current problem setting, the only (partially) observed attribute is identity. Having additional information would be useful for further improving the classification accuracy, and could be easily added to DS PSL. Extra information can include network structure, such as information about an individual's social circles [17], or prior estimates about an individual's preferences based on other attributes [25]. Second, we believe that DS PSL and DS-PC PSL can be extended to other attribute prediction settings. As discussed in Section 3.2, the concept of separating identity and preference is general and is not limited to social networks, and we can not only decouple identity from preference, but also other attributes associated with individuals. Third, it would be interesting to evaluate DS PSL on datasets with varying degrees of homogeneity. Our empirical results demonstrates that DS PSL can

work relatively well in a setting where homophily in identity is weak. Ultimately, we would like to build a hybrid model which can automatically detect the homogeneity properties of the graph, and adjust the weight between homophily and decoupled smoothing accordingly.

8 ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation grants CCF-1740850 and IIS-1703331, AFRL and the Defense Advanced Research Projects Agency. We also thank the reviewers for their constructive feedback that helped shape the paper.

REFERENCES

[1] Kristen M. Altenburger and Johan Ugander. 2018. Monophily in social networks introduces similarity among friends-of-friends. *Nature Human Behaviour* 2, 4 (2018), 284–290.
[2] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *Journal of Machine Learning Research (JMLR)* 18, 1 (2017), 3846–3912.
[3] Smriti Bhagat, Graham Cormode, and S. Muthukrishnan. 2011. *Node Classification in Social Networks.* Springer, Boston, MA, USA, 115–148.
[4] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. 2007. Combining Collective Classification and Link Prediction. In *Workshop on Mining Graphs and Complex Structures at the IEEE International Conference on Data Mining (ICDM).* IEEE, Omaha, NE, USA.
[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* 3, 1 (2011), 1–122.
[6] Ivan Brugere, Brian Gallagher, and Tanya Y. Berger-Wolf. 2018. Network Structure Inference, A Survey: Motivations, Methods, and Applications. *ACM Computing Survey* 51, 2 (2018), 1–39.
[7] Alex Chin, Yatong Chen, Kristen M. Altenburger, and Johan Ugander. 2019. Decoupled Smoothing on Graphs. In *The World Wide Web Conference (WWW).* ACM, San Francisco, CA, USA.
[8] Ming Ding, Jie Tang, and Jie Zhang. 2018. Semi-Supervised Learning on Graphs with Generative Adversarial Nets. In *ACM International Conference on Information and Knowledge Management (CIKM).* ACM, Torino, Italy.
[9] Dhivya Eswaran, Srijan Kumar, and Christos Faloutsos. 2020. Higher-Order Label Homogeneity and Spreading in Graphs. In *The World Wide Web Conference (WWW).* ACM, New York, NY, USA.
[10] Brian Gallagher, Hanghang Tong, Tina Eliassi-Rad, and Christos Faloutsos. 2008. Using Ghost Edges for Classification in Sparsely Labeled Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD).* ACM, New York, NY, USA.
[11] Nadia Ghamrawi and Andrew McCallum. 2005. Collective Multi-Label Classification. In *ACM International Conference on Information and Knowledge Management (CIKM).* ACM, Bremen, Germany.
[12] Aditya Grover, Aaron Zweig, and Stefano Ermon. 2019. Graphite: Iterative Generative Modeling of Graphs. In *International Conference on Machine Learning (ICML).* PMLR, Long Beach, CA, USA.
[13] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation Learning on Graphs: Methods and Applications. *IEEE Data Engineering Bulletin* 40, 3 (2017), 52–74.
[14] Paul F. Lazarsfeld and Robert K. Merton. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society* 18, 1 (1954), 18–66.
[15] Qing Lu and Lise Getoor. 2003. Link-Based Classification. In *International Conference on Machine Learning (ICML).* AAAI, Washington, DC, USA.
[16] Sofus A. Macskassy and Foster Provost. 2003. A Simple Relational Classifier. In *Workshop on Multi-Relational Data Mining (MRDM).* ACM, Washington, DC, USA.
[17] Julian Mcauley and Jure Leskovec. 2014. Discovering Social Circles in Ego Networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 1–28.
[18] Luke K McDowell and David W Aha. 2012. Semi-Supervised Collective Classification with Hybrid Label Regularization. In *International Conference on Machine Learning (ICML).* Omnipress, Edinburgh, Scotland.
[19] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.
[20] John Moore and Jennifer Neville. 2017. Deep Collective Inference. In *The AAAI Conference on Artificial Intelligence (AAAI).* AAAI, San Francisco, CA, USA.
[21] Jennifer Neville and David Jensen. 2000. Iterative Classification in Relational Data. In *AAAI Workshop on Learning Statistical Models From Relational Data (SRL).* AAAI, Austin, TX, USA.
[22] Mark. E. J. Newman, Duncan J. Watts, and Steven H. Strogatz. 2002. Random graph models of social networks. *Proceedings of the National Academy of Sciences (PNAS)* 99, suppl 1 (2002), 2566–2572.
[23] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine* 29, 3 (2008), 93–106.
[24] Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. 2012. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 16 (2012), 4165–4180.
[25] Kamil Wais. 2016. Gender Prediction Methods Based on First Names with genderizeR. *The R Journal* 8, 1 (2016), 17–37.
[26] Stanley Wasserman and Katherine Faust. 1994. *Triads.* Vol. 8. Cambridge University Press, Cambridge, England, 556–602.
[27] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440–442.
[28] Dengyong Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with Local and Global Consistency. In *Advances in Neural Information Processing Systems (NIPS).* MIT, Cambridge, MA, USA.
[29] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *International Conference on Machine Learning (ICML).* AAAI, Washington, DC, USA.