Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data

Victor Veitch* **Morgane Austern*** **Wenda Zhou*** **David M. Blei** **Peter Orbanz**

Abstract

Empirical risk minimization is the principal tool for prediction problems, but its extension to relational data remains unsolved. We solve this problem using recent advances in graph sampling theory. We (i) define an empirical risk for relational data and (ii) obtain stochastic gradients for this risk that are automatically unbiased. The key ingredient is to consider the method by which data is sampled from a graph as an explicit component of model design. Theoretical results establish that the choice of sampling scheme is critical. By integrating fast implementations of graph sampling schemes with standard automatic differentiation tools, we are able to solve the risk minimization in a plug-and-play fashion even on large datasets. We demonstrate empirically that relational ERM models achieve state-of-the-art results on semi-supervised node classification tasks. The experiments also confirm the importance of the choice of sampling scheme.

1 Introduction

Relational data is data that can be represented as a graph (e.g. the link graph of a social network, or user/movie ratings), possibly annotated with additional information (e.g. user profiles). We consider prediction problems for relational data. Many prediction methods used in machine learning are based on empirical risk minimization (ERM) [17, 18, 15]. However, ERM inherently assumes that data are i.i.d. This assumption is meaningless for relational data; hence, generalizing ERM to relational data remains an unsolved problem.

To address this problem, we draw on recent work in statistics and applied probability that emphasizes the role of sampling theory in modeling graph data [12, 19, 2, 3]. In Section 2, we explain how risk and empirical risk can be defined for relational data. The definition is based on a specific choice of a randomized algorithm that samples data from a graph. Different sampling algorithms yield different notions of relational empirical risk, and we review several possible choices. For large datasets, the classical ERM problem is usually solved with stochastic gradient descent (SGD), since unbiased estimates of the empirical risk can be efficiently computed by uniformly subsampling the data. In Section 2.2, we show that unbiased stochastic gradients of the relational empirical risk can be efficiently computed by replacing uniform sampling with an appropriate surrogate. This results in an efficient plug-and-play SGD algorithm for solving the minimization problem. In Section 3, we detail two concrete examples, and we draw connections to the graph embeddings literature [e.g 13, 6, 16, 8]. We derive theoretical results in Section 4 that clarify the fundamental role of the choice of sampling algorithm. In Section 5, we study relational ERM empirically. We observe that (i) the choice of sampling scheme has a substantial effect in practice, (ii) novel models can be easily fit using SGD with relational ERM, and (iii) combining these two observations leads to state-of-the-art performance on semi-supervised vertex label prediction tasks. We provide fast implementations of a variety of graph sampling algorithms and integration with TensorFlow.[2]

*equal contribution

[2] github.com/wooden-spoon/relational-ERM

Preprint. Work in progress.

2 Empirical risk and subsampling

Here is the classical setup of ERM: A single observation is represented by a random variable X. Associated with X is a label Y, which may or may not be observed. We denote the "complete" observation $\overline{X} = (X, Y)$. A **predictor** is a function π that maps observations to labels. For an element x of the sample space, it is typically written as $\pi(x) = \hat{y}$, where $\hat{y}$ is a predicted label. Anticipating relational data, we instead write $\hat{x} = (x, \hat{y})$ and $\pi(x) = \hat{x}$, that is, the predictor "completes" unlabeled information x to labeled information $\hat{x}$. How well $\hat{x}$ reconstructs $\overline{x}$ is measured by the value $L(\hat{x}, \bar{x})$ of a **loss function** L with values in $[0, \infty]$.

ERM typically assumes a sample $\overline{\mathbb{S}}_n = (\overline{X}_1, \ldots, \overline{X}_n)$ of observations to be generated i.i.d. from some probability distribution P. The **risk** of π is the expectation of the loss under this distribution,

$$R(\pi) \;:=\; \mathbb{E}_{\overline{X}\sim P}[L(\pi(X), \overline{X})] \;. \tag{1}$$

The **empirical distribution** of the sample $\mathbb{S}_n$ is $\mathbb{F}(\mathbb{S}_n) := \frac{1}{n}\sum_{i\leq n} \delta_{\overline{X}_i}$, where $\delta_{\bar{x}}$ is either the Dirac function at $\bar{x}$ (if the set in which $\bar{x}$ takes its values is uncountable), or the indicator $\delta_{\bar{x}}(\,\bullet\,) = \mathbb{1}\{\bullet = \bar{x}\}$ (on a discrete space). The **empirical risk** of π substitutes $\mathbb{F}$ for P,

$$\hat{R}(\pi, \overline{\mathbb{S}}_n) \;:=\; \mathbb{E}_{\overline{X}\sim\mathbb{F}(\overline{\mathbb{S}}_n)}[L(\pi(X), \overline{X})|\overline{\mathbb{S}}_n] \;=\; \frac{1}{n}\sum_{i\leq n} L(\pi(X_i), \overline{X}_i) \;. \tag{2}$$

Since the observed sample is random, $\hat{R}(\pi, \mathbb{S}_n)$ is random. To choose a predictor π, we posit a hypothesis class $\{\pi_\theta | \theta \in \mathcal{T}\}$ of predictors indexed by a parameter θ with values in a parameter space $\mathcal{T}$. **Empirical risk minimization** selects a predictor $\hat{\pi}$ as

$$\hat{\pi} := \pi_{\hat{\theta}_n} \qquad \text{where} \qquad \hat{\theta}_n \;:=\; \operatorname*{argmin}_\theta \hat{R}(\pi_\theta, \mathbb{S}_n) \;.$$

ERM is statistically sound in that $\hat{\theta}_n$ converges to a value θ^* determined by P. This property follows from the fact that $X_1, \ldots, X_n$ are i.i.d. [4].

2.1 Empirical risk for relational data

Classical ERM underpins many machine learning algorithms, but it tacitly assumes data are i.i.d. from some distribution. For relational data, the i.i.d. assumption is meaningless. We now consider how to define a useful notion of empirical risk for relational data.

We model relational data as graphs. Instead of the sample $\overline{\mathbb{S}}_n$ above, we observe a graph $\overline{G}_n$ of size n (for example, the number of vertices or edges), possibly annotated, e.g., by vertex labels. This graph is assumed to represent a small part of a large, underlying graph $\mathcal{G}$ (e.g., an entire social network). For relational data, the predictor π is a function whose input is an incomplete version G_n of $\overline{G}_n$, which it augments with an estimate of the missing information, such as vertex labels or missing edges.

To generalize the empirical risk, we first re-examine its definition above: Suppose the observations $\overline{X}_i$ are generated by randomly selecting examples from some large, underlying set $\mathcal{X}$. In statistics, $\mathcal{X}$ is called a **population**. We generate a sample as follows:

Algorithm 1 (Sampling with replacement)**.**
i.) Select n elements $\overline{X}_i$ of $\mathcal{X}$ independently and uniformly with replacement.
ii.) Report $\overline{\mathbb{S}}_n = (\overline{X}_1, \ldots, \overline{X}_n)$.

Then $\overline{X}_1, \ldots, \overline{X}_n$ are i.i.d., with distribution P determined by $\mathcal{X}$. If $\mathcal{X}$ is small, one has to distinguish carefully between sampling with and without replacement, but for large populations, the two become indistinguishable. In the so-called *infinite population limit* $|\mathcal{X}| \to \infty$, every distribution P arises in this manner. Modeling data as i.i.d. can thus be justified as sampling from a very large population using Algorithm 1. The empirical risk (2) and risk (1) can thus be rewritten as

$$\hat{R}(\pi, \overline{\mathbb{S}}_n) = \mathbb{E}_{\overline{\mathbb{S}}_1\sim\overline{\mathbb{S}}_n}[L(\pi(\mathbb{S}_1), \overline{\mathbb{S}}_1)] \quad \text{and} \quad R(\pi) = \mathbb{E}_{\overline{\mathbb{S}}_1\sim\mathcal{X}}[L(\pi(\mathbb{S}_1), \overline{\mathbb{S}}_1)] = \mathbb{E}_{\overline{\mathbb{S}}_n\sim\mathcal{X}}[\hat{R}(\pi, \overline{\mathbb{S}}_n)] \;.$$

See e.g. [14] for a rigorous justification. Note that the empirical risk $\hat{R}$ involves two sampling steps: A data acquisition step $\overline{\mathbb{S}}_n \sim \mathcal{X}$ that generates an observed sample of data by drawing from the (unobserved) underlying population, and a draw from $\overline{\mathbb{S}}_1 \sim \overline{\mathbb{S}}_n$ from the empirical measure.

We use this perspective on classical ERM to define ERM for relational data. In this case, the population $\mathcal{X}$ is replaced by the population graph $\mathcal{G}$. Again, there are two distributions at play. In the data acquisition step, we replace Algorithm 1 by a suitable algorithm $\mathbb{A}$ that samples a random subgraph $\overline{G}_n \sim_{\mathbb{A}} \mathcal{G}$ from an input graph $\mathcal{G}$. An observed graph $\overline{G}_n$ is then modeled as such a sample from a large, unknown population graph $\mathcal{G}$. Specific choices for $\mathbb{A}$ are discussed in Section 2.3 below. As in the i.i.d. case above, an infinite-population limit $|\mathcal{G}| \to \infty$ can be made rigorous [12, 2].

The second sampling step—which corresponds to sampling from the empirical distribution in the i.i.d. case—uses sampling algorithm $\mathbb{B}$, which may or may not be identical to $\mathbb{A}$. We denote this second sampling step as a subsampling routine

$$\mathsf{Sample}(\overline{G}_n, k) := \overline{G}_k \sim_{\mathbb{B}} \overline{G}_n.$$

The distinction in notation is made to emphasize that $\overline{G}_n \sim_{\mathbb{A}} \mathcal{G}$ is a *modeling assumption* on how observed data has been acquired, whereas Sample is a *definition*, chosen by the analyst and implemented in code. In analogy to the sampling representation of the risk above, we define

$$\hat{R}_k(\pi, \overline{G}_n) := \mathbb{E}_{\overline{G}_k = \mathsf{Sample}(\overline{G}_n, k)}[L(\pi(G_k), \overline{G}_k) \mid \overline{G}_n] \quad \text{and} \quad R_k(\pi) = \mathbb{E}_{\overline{G}_n \sim \mathcal{G}}[\hat{R}_k(\pi, \overline{G}_n)] \,. \tag{3}$$

We call R_k the **relational risk**, and $\hat{R}_k$ the **relational empirical risk**. **Relational empirical risk minimization** selects a predictor $\hat{\pi}$ as

$$\hat{\pi} := \pi_{\hat{\theta}_n} \quad \text{where} \quad \hat{\theta}_n := \operatorname*{argmin}_{\theta} \hat{R}_k(\pi_\theta, \overline{G}_n) \,. \tag{4}$$

In summary, a **relational ERM model** is defined by three ingredients:

1. A class of predictors $\{\pi_\theta | \theta \in \mathcal{T}\}$ with parameter θ.
2. A loss function L.
3. A sampling routine Sample.

Given observed data $\overline{G}$, a model is fit by solving (4).

2.2 Stochastic gradient descent

For relational ERM to be useful in practice, the minimization problem (4) must be (approximately) solvable in a reasonable amount of time. This is possible through stochastic gradient descent (SGD). SGD requires unbiased estimates of the gradient of the objective function—in this case, the relational empirical risk. We define a stochastic gradient as $\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta)$, the gradient of the loss computed on a sample of size k drawn with Sample. The key observation is

$$\nabla_\theta \hat{R}_r(\theta; G_n) \;=\; \nabla_\theta \mathbb{E}[L(\mathsf{Sample}(G_n, k), \theta) \mid G_n] \;=\; \mathbb{E}[\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta) \mid G_n].$$

That is, the random gradient $\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta)$ is an unbiased estimator of the gradient of the full relational empirical risk. If Sample is computationally efficient, then relational ERM can be solved using SGD with this stochastic estimator. Combined with automatic differentiation, the resulting algorithm is an effective and generically applicable solver for the minimization problem.

2.3 Subsampling algorithms

In classical ERM, sampling uniformly (with or without replacement) is typically the only choice. In contrast, there are many ways to sample from a graph. Each such sampling algorithm Sample leads to a different notion of risk and empirical risk in (3). This section describes some possibilities.

Perhaps the closest analogue of Algorithm 1 for graphs is uniform selection of vertices:

Algorithm 2 (Unifom vertex sampling)**.**
- i.) Select n vertices of g independently and uniformly without replacement.
- ii.) Extract the induced subgraph $\overline{G}_n$ of g on these vertices.
- iii.) Label the vertices of $\overline{G}_n$ by $1, \ldots, k$ in order of appearance.

The input graph g may either represent a population $\mathcal{G}$, or a previously extracted sample G_n. Algorithm 2 is simple, but often problematic, since it is not suitable for sparse graphs. There are various definitions of sparsity, but they all have in common that the fraction $\rho = {}^{e(G_n)}/_{v(G_n)^2}$ of edges present in the graph approaches zero as the graph grows. The expected number of edges reported by Algorithm 2 is $k^2\rho$, which vanishes for large graphs. Algorithms for sparse data must mitigate this problem. The next algorithm retains only non-empty portions of the graph:

Algorithm 3 (p-sampling [19])**.**
- i.) Select each vertex in g independently, with a fixed probability $p \in [0, 1]$.
- ii.) Extract the induced subgraph $\overline{G}_n$ of g on the selected vertices.
- iii.) Delete all isolated vertices from $\overline{G}_n$, and report the resulting graph.

The main difference to Algorithm 2 is the deletion step (iii).

Algorithm 4 (Uniform edge sampling)**.**
- i.) Select n edges in g uniformly and independently from the edge set.
- ii.) Report the graph $\overline{G}_n$ consisting of these edges, and all vertices incident to these edges.

Recall that a **simple random walk** of length k on a graph g selects vertices $v_1, \ldots, v_k$ by starting at a given vertex v_1, and drawing each vertex v_{i+1} uniformly from the neighbors of v_i. Typically, a random walk sample is augmented with additional edges, either as part of the data collection procedure, or to increase the efficiency of stochastic gradient descent. We consider two strategies. The first fills in the entire induced subgraph:

Algorithm 5 (Random walk: Induced)**.**
- i.) Sample a random walk $v_1, \ldots, v_k$ starting at a uniformly selected vertex of g.
- ii.) Report $\overline{G}_n$ as the edge list of the vertex induced subgraph of the walk.

The second strategy augments the walk by hallucinating plausible additional edges.

Algorithm 6 (Random walk: Skipgram [13])**.**
- i.) Sample a random walk $v_1, \ldots, v_k$ starting at a uniformly selected vertex of g.
- ii.) Report $\overline{G}_n = \{(v_i, v_j) : d(v_i, v_j) < W\}$, where the *window* W is a sampler parameter, and $d(v_i, v_j)$ is the number of steps between v_i and v_j.

This algorithm relates to the Skipgram model [11], and is commonly used in graph embeddings, e.g. DeepWalk [13] and its successors.

Remark 2.1. These algorithms presuppose graph data, but all can be generalized to "bipartite" relational data, where the rows and columns of an input matrix represent different entities (for example, users and movies). In this case, Algorithm 2 would select n_r rows and n_c columns, Algorithm 3 would use two parameters p_r and p_c, Algorithm 4 would still select n entries of the matrix and report each along with a row and column identifier. Random walks can also be transcribed to this setting, although there is not a unique way of doing so.

2.4 Negative sampling

For a pair of vertices in an input graph g, a sampling algorithm can report three types of edge information: The edge may be observed as present, observed as absent (a *non-edge*), or may not be observed Aside from Algorithm 2, the algorithms above do not treat edge and non-edge information equally: Algorithms 4 to 6 cannot report non-edges, and the deletion step in Algorithm 3 biases it towards edges over non-edges. However, the locations of non-edges can carry significant information.

Negative sampling schemes are "add-on" algorithms that are applied to the output of a graph sampling algorithm and augment it by non-edge information. Let $\overline{G}_n$ denote a sample generated by one of the algorithms above from an input graph g.

Algorithm A (Negative sampling: Induced subgraph)**.**
- i.) Report the subgraph induced by $\overline{G}_n$, in the input graph g from which $\overline{G}_n$ was drawn.

Another method [10, 5] is based on the unigram distribution: Define a probability distribution on the vertex set of g by $P_n(v) := \text{Prob}\{v \in \overline{H}_n\}$, the probability that v would occur in a separate, independent sample $\overline{H}_n$ generated from g by the same algorithm as $\overline{G}_n$. For $\tau > 0$, we define a distribution $P_n^\tau(v) := (P_n(v))^\tau / Z(\tau)$, where $Z(\tau)$ is the appropriate normalization.

Algorithm B (Negative sampling: Unigram)**.**
For each vertex v in $\overline{G}_n$:
i.) Select k vertices $v_1, \ldots, v_k \sim P_n^\tau$ independently.
ii.) If (v, v_j) is a non-edge in g, add it to $\overline{G}_n$.

Algorithm B is common in the embeddings literature, where the canonical choice is $\tau = \frac{3}{4}$, see [10].

3 Example Models

We now consider two concrete examples of relational ERM, and also discuss relations to graph embedding methods. For the examples, we specify a class of predictors π_θ and a loss function L. Each can then be used with different choices of Sample. In both examples, we observe a natural split of the parameter into a pair $\theta = (\gamma, \lambda)$: global parameters γ shared across the entire graph, and embedding parameters λ, where each vertex v has an associated embedding λ_v. Informally, global parameters encode population properties—"people with different political affiliation are less likely to be friends"—and the embeddings encode vertex-specific information—"Bob is a radical vegan."

Semi-supervised node classification Consider a network G_n where each node i is labeled by some binary features—for example, hyperlinked documents labeled by subjects, or interacting proteins labeled by function. The task is to predict the features of a subset of these nodes given the graph structure and the features of the remaining nodes.

The model has the following form: Each vertex i is assigned a k-dimensional embedding vector $\lambda_i \in \mathbb{R}^k$. Feature prediction is made by a parameterized function $f(\,\bullet\,; \gamma) : \mathbb{R}^k \to [0,1]^L$ that maps the embedding of each node to the probability that the node has each of the possible features. Let σ denote the sigmoid function; let $l_{ij} \in \{0, 1\}$ denote whether vertex i has feature j; and let $q \in [0, 1]$. The loss on subgraphs $G_k \subset G_n$ is:

$$\begin{aligned} L(G_k, l; \lambda, \gamma) \;&=\; q\Big(\sum_{i \in v(G_k)} \sum_{j=1}^{L} l_{ij} \log f(\lambda_i; \gamma)_j + (1 - l_{ij}) \log(1 - f(\lambda_i; \gamma)_j) \Big) \\ &+\; (1-q)\Big(- \sum_{i,j \in e(G_k)} \log \sigma(\lambda_j^T \lambda_i) \;-\; \sum_{i,j \in \bar{e}(G_k)} \log(1 - \sigma(\lambda_j^T \lambda_i)) \Big) \,. \end{aligned} \tag{5}$$

Here, v, e, and $\bar{e}$ denote the vertices, edges, and non-edges of the graph respectively. The loss on edge terms is cross-entropy, a standard choice in embedding models [8]. Intuitively, the predictor uses the embeddings to predict both the vertex labels and the subgraph structure.

The model is completed by choosing a sampling scheme Sample. Relational ERM then fits the parameters as

$$(\hat{\lambda}_n, \hat{\gamma}_n) \;=\; \operatorname*{argmin}_{\lambda, \gamma} \mathbb{E}[L(\mathsf{Sample}(G_n, k), l; \lambda, \gamma) \mid G_n].$$

In Section 5 below, we revisit this model and report prediction performance for different choices of Sample.

Wikipedia category embeddings We consider Wikipedia articles joined by hyperlinks. Each article is tagged as a member of one or more categories—for example, "Muscles_of_the_head_and_neck", "Japanese_rock_music_groups", or "People_from_Worcester." The task is to learn latent structure encoding semantic relationships between the categories.

Let G_n denote the hyperlink graph and let $\mathcal{C}(i)$ denote the categories of article i. Each category $c \in C$ is assigned an embedding γ_c, and the embedding of each article (vertex) is taken to be the sum

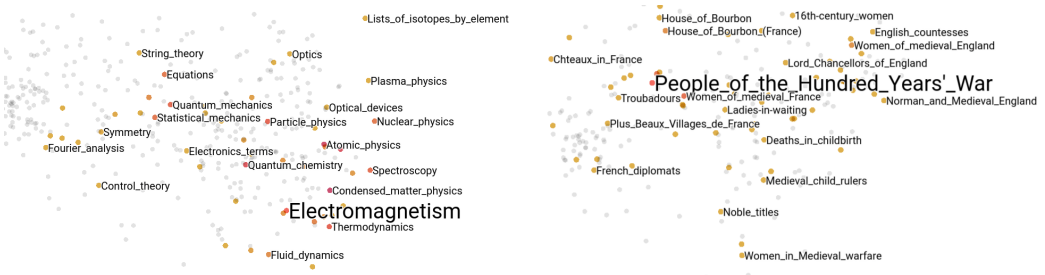

Figure 1: Trained Wikipedia category embeddings. *Left*: Categories that are nearest neighbors of "Electromagnetism", projected to maximally separate "Mathematics" and "Physics" on the horizontal axis. *Right*: Neighbors of "People_of_the_hundred_years'_war", projected to separate "France" and "England".

of the embeddings of its categories, $\lambda_i := \sum_{c \in \mathcal{C}(i)} \gamma_c$. The loss is

$$L(G_k, C; \lambda) = - \sum_{i,j \in e(G_k)} \log \sigma(\lambda_j^T \lambda_i) \; - \sum_{i,j \in \bar{e}(G_k)} \log(1 - \sigma(\lambda_j^T \lambda_i)) \; , \tag{6}$$

where e and $\bar{e}$ denote, respectively, the presence and absence of hyperlinks between articles. Intuitively, the predictor uses the category embeddings to predict the hyperlink structure of subgraphs. Relational ERM chooses the embeddings as

$$\hat{\gamma}_n = \operatorname*{argmin}_{\gamma} \mathbb{E}[L(\mathsf{Sample}(G_n, k), C; \lambda(\gamma)) \mid G_n] \; .$$

We write $\lambda(\gamma)$ to emphasize that the article embeddings are a function of the category embeddings. Category embeddings obtained with this model are illustrated in Figure 1; see Section 5 for details on the experiment.

Graph representation learning Methods for learning embeddings of vertices are widely studied; see Hamilton, Ying, and Leskovec [8] for a recent review. The prototypical graph representation learning algorithm is DeepWalk [13]. The basic approach is to draw a large collection of simple random walks, view each of these walks as a "sentence" where each vertex is a "word", and learn vertex embeddings by a applying a standard word embedding method [11, 10]. This idea has been extended a number of directions, including using more complicated random walks [6], incorporating covariate information [7], developing GANs for graphs [1]. These approaches are state of the art for many tasks.

Graph representation learning algorithms may be viewed as particular cases of relational ERM. For example, informally, DeepWalk is equivalent to a relational ERM model that (i) predicts graph structure using a predictor parameterized only by embedding vectors, (ii) uses a cross entropy loss on graph structure, and (iii) samples by the random-walk skipgram sampler with unigram negative sampling.[3] The relational ERM perspective allows us to move beyond random walk based sampling. It also permits modifications such as including covariate information or training embeddings and labels simultaneously.

4 Theory

Relational ERM involves two sampling procedures: The algorithm $\mathbb{A}$ used in the data acquisition step $\overline{G}_n \sim_{\mathbb{A}} \mathcal{G}$, which generates data from an underlying population $\mathcal{G}$, and the algorithm Sample used to defined ERM and to execute SGD. Since Sample is a component of model design, different choices of Sample should lead to different results, even in the infinite-data limit. The results below show that this is the case.

To phrase a theoretical result, we must choose specific algorithms. For data acquisition, we consider a population graph $\mathcal{G}$ with $|\mathcal{G}|$ edges. We assume that $\mathcal{G}$ is "very large," in the sense that $|\mathcal{G}| \to \infty$.

[3] DeepWalk uses hierarchical softmax instead of negative sampling, so the correspondence is not literal.

In other words, any effect that weakens as the graph grows is assumed to be negligible. We assume that an observed sample $\overline{G}_n$ of size n is generated by p-sampling from $\mathcal{G}$, with $p = n/\sqrt{|\mathcal{G}|}$, for $|\mathcal{G}| \to \infty$. The distribution of $\overline{G}_n$ in the "infinite population" case $|\mathcal{G}| \to \infty$ is well-defined [2].

Under this modeling assumption on $\mathbb{A}$, we compare two choices of $\mathsf{Sample}(G_n, k)$: One is p-sampling with $p = k/\sqrt{n}$. The other is sampling using a simple random walk (Algorithm 5), with walk length k. We denote the empirical risk Algorithm 3 defines by $\hat{R}_k^{\mathrm{ps}}$, and of Algorithm 5 by $\hat{R}_k^{\mathrm{rw}}$.

The first result establishes that the relational empirical risk is a sensible objective function, and that the trained model depends on Sample even in the infinite-data limit.

Theorem 4.1. *Suppose that G_n is collected by p-sampling as described above. Further suppose technical conditions given in the appendix. Then, for parameter setting $\bar{\theta}$ satisfying the technical conditions, there are constants $c_{\bar{\theta}}^{\mathrm{ps}}, c_{\bar{\theta}}^{\mathrm{rw}} \in \mathbb{R}_+$ such that*

$$\hat{R}_k^{\mathrm{ps}}(\bar{\theta}; \overline{G}_n) \to c_{\bar{\theta}}^{\mathrm{ps}} \qquad \hat{R}_k^{\mathrm{rw}}(\bar{\theta}; \overline{G}_n) \to c_{\bar{\theta}}^{\mathrm{rw}} \tag{7}$$

both a.s. and in L_1 as $n \to \infty$. Moreover, there are constants $c_*^{\mathrm{ps}}, c_*^{\mathrm{rw}} \in \mathbb{R}_+$ such that

$$\min_\theta \hat{R}_k^{\mathrm{ps}}(\theta; \overline{G}_n) \to c_*^{\mathrm{ps}} \qquad \min_\theta \hat{R}_k^{\mathrm{rw}}(\theta; \overline{G}_n) \to c_*^{\mathrm{rw}}, \tag{8}$$

both a.s. and in L_1, as $s \to \infty$.

More details regarding the constants in (7) are given in the appendix; the relevant property is that $c_{\bar{\theta}}^{\mathrm{ps}}$ and $c_{\bar{\theta}}^{\mathrm{rw}}$ are generally distinct. Thus, (7) states that the limits of the p-sampling empirical risk and random-walk empirical risk will not agree because the learning procedure depends on the choice of Sample. The same observation holds, as (8) shows, for the minimal empirical risk determined by learning.

The next result strengthens the convergence guarantee. It shows that the estimated global parameters converge in an absolute sense as more data is collected.

Theorem 4.2. *Suppose the conditions of Theorem 4.1, and also that the loss function verifies a certain strict convexity property in γ, given explicitly in the appendix. Let $\tilde{\gamma}_n^{\mathrm{ps}} = \operatorname{argmin}_\gamma \min_\lambda \hat{R}_k^{\mathrm{ps}}(\gamma, \lambda; \overline{G}_n)$, and similarly for $\tilde{\gamma}_n^{\mathrm{rw}}$. Then $\tilde{\gamma}_n^{\mathrm{ps}} \to \tilde{\gamma}_*^{\mathrm{ps}}$ and $\tilde{\gamma}_n^{\mathrm{rw}} \to \tilde{\gamma}_*^{\mathrm{rw}}$ almost surely for some constants $\tilde{\gamma}_*^{\mathrm{ps}}$ and $\tilde{\gamma}_*^{\mathrm{rw}}$.*

Again, $\tilde{\gamma}_*^{\mathrm{ps}} \neq \tilde{\gamma}_*^{\mathrm{rw}}$ in general, and the choice of Sample hence affects learned parameters even in the infinite data limit. In general, $\tilde{\gamma}_n^{\mathrm{ps}}$ need not coincide with the global parameter estimate $\hat{\gamma}_n^{\mathrm{ps}}$ learned by simultaneously minimizing the global parameters and embeddings. However, because the parameter values learned by simultaneous minimization need not be identifiable, it seems necessary to consider the two stage procedure to establish a simple convergence result.

5 Experiments

We empirically study the example models in Section 3, defined by (5) and (6) respectively.[4] The models are determined by (5) and (6) up to the choice of Sample. The experiments (i) consider the influence of the choice of sampling scheme; (ii) illustrate its applicability to new tasks; and (iii) evaluate performance.

Node classification problems We begin with the semi-supervised node classification task described in Section 3, using the model (5) with different choices of Sample. We study the blog catalog and protein-protein interaction data reported in [6], summarized by the table on the right. Each vertex in the graph has one or more labels, and 50% have their labels censored at training time. The task is to predict these labels at test time.

	Blogs	Protein
Vertices	10,312	3,890
Edges	333,983	76,584
Feature Dim.	39	50

Two-stage training. We first train the model (5) using no label information to learn the embeddings (that is, with $q = 0$). We then fit a logistic regression to predict vertex features from

[4] Code at github.com/wooden-spoon/relational-ERM

the trained embeddings. This two stage approach is a standard testing procedure in the graph embedding literature, e.g. [13, 6]. We use the same scoring procedure as Node2Vec [6] and, where applicable, the same hyperparameters. We preprocess the data to remove self-edges, and restrict each network to the largest connected component. The table on the right shows the effect of varying the sampling scheme used to train the embeddings. SGD succeeds in solving the relational ERM problem for all sampling schemes. As expected, we observe that the choice of sampling scheme affects the embeddings produced via the learning procedure, and thus also the outcome of the experiment. We further observe that sampling non-edges by unigram negative sampling gives better predictive performance relative to selecting non-edges from the vertex induced subgraph.

Simultaneous training. Next, we fit the model of Section 3 with $q = 0.001$—training the embeddings and global variables simultaneously. We choose label predictor π as logistic regression, and adapt the loss to measure the loss only on vertices in the positive sample. We report average macro F1 scores, computed using the node2vec scoring procedure:

Choice of Sample	Alg. #	Blogs	Protein
p-samp+ind.	3+A	0.17	0.14
p-samp+ns	3+B	0.22	0.16
rw/skipgram+ns	6+B	0.18	0.16
rw/induced+ind	5+A	0.08	0.08
rw/induced+ns	5+B	0.18	0.16
unif. edge+ns	2+B	0.21	0.15

	Blog catalog			Protein-Protein		
ERM defined by	Unif.	p-samp	rw	Unif.	p-samp	rw
p-samp+ns (Alg. 3+B)	0.30	0.34	0.35	0.30	0.37	0.39
rw/skipgram+ns (Alg. 6+B)	0.20	0.26	0.27	0.25	0.32	0.34
Node2Vec (reported)	0.26	-	-	0.18	-	-

Columns are labeled by the sampling scheme used to draw test vertices. We observe:

- Learning embeddings and logistic regression simultaneously improves performance.
- When training jointly, p-sampling outperforms the standard rw/skipgram procedure.
- Labels of nodes selected by random walk or p-sampling are easier to predict than those chosen uniformly at random.

Note that the average computed with uniform vertex sampling is the standard scoring procedure used in the previous table.

Wikipedia Category Embeddings Finally, we illustrate an application of relational ERM to a non-standard task. We consider the task of discovering semantic relations between Wikipedia categories, as described in Section 3. We define a relational ERM model by choosing the cost function L in (6), and Sample as 6+B, the skipgram random walk sampler with unigram negative sampling. The data $\overline{G}_n$ is the Wikipedia hyperlink network from [9], consisting of Wikipedia articles from 2011-09-01 restricted to articles in categories containing at least 100 articles. The dataset is relatively large—about 1.8M nodes and 28M edges. We choose embedding dimension $k = 128$. SGD converges in about 90 minutes on a desktop computer equipped with a Nvidia Titan Xp GPU. Figure 1 on page 6 visualizes example trained embeddings.

6 Conclusion

Relational ERM is a generalization of ERM from i.i.d. data to relational data. The key ingredients are modeling the sampling scheme by which the data is collected—the analogue for the i.i.d. assumption—and explicitly specifying the sampling scheme used to subsample the data—the analogue of the empirical distribution. Relational ERM models are defined by a loss function, a predictor class, and a sampling scheme. The models can then be fit automatically using SGD. Accordingly, relational ERM provides an easy method to specify and fit relational data models, as illustrated in Sections 3 and 5.

The results presented here suggest a number of directions for future inquiry. Foremost: what is the relational analogue of statistical learning theory? The theory derived in the present paper establishes initial results. A more complete treatment may provide statistical guidelines for model development. Our results hinge critically on the assumption that the data is collected by p-sampling; it is natural to

ask whether other data-generating mechanisms can be accommodated. Similarly, it is natural to ask for guidelines for the choice of Sample.

References

[1] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann. *NetGAN: Generating Graphs via Random Walks*. 2018. arXiv: 1803.00816.

[2] C. Borgs, J. T. Chayes, H. Cohn, and V. Veitch. *Sampling perspectives on sparse exchangeable graphs*. 2017. arXiv: 1708.03237.

[3] H. Crane and W. Dempsey. *A framework for statistical network modeling*. 2015. arXiv: 1509.08185.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

[5] Y. Goldberg and O. Levy. *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. 2014. arXiv: 1402.3722.

[6] A. Grover and J. Leskovec. "Node2Vec: Scalable Feature Learning for Networks". In: *Proc. 22nd Int. Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 2016, pp. 855–864.

[7] W. L. Hamilton, R. Ying, and J. Leskovec. *Inductive Representation Learning on Large Graphs*. June 2017. arXiv: 1706.02216.

[8] W. L. Hamilton, R. Ying, and J. Leskovec. *Representation Learning on Graphs: Methods and Applications*. 2017. arXiv: 1709.05584.

[9] C. Klymko, D. Gleich, and T. G. Kolda. *Using Triangles to Improve Community Detection in Directed Networks*. 2014. arXiv: 1404.5874.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781.

[12] P. Orbanz. *Subsampling large graphs and invariance in networks*. 2017. arXiv: 1710.04217.

[13] B. Perozzi, R. Al-Rfou, and S. Skiena. "DeepWalk: Online Learning of Social Representations". In: *Proc. 20th Int. Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 2014, pp. 701–710.

[14] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

[15] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[16] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. "LINE: Large-scale Information Network Embedding". In: *Proc. 24th Int. Conference on World Wide Web (WWW '15)*. 2015, pp. 1067–1077.

[17] V. Vapnik. "Principles of Risk Minimization for Learning Theory". In: *Advances in Neural Information Processing Systems 4*. 1992, pp. 831–838.

[18] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[19] V. Veitch and D. M. Roy. "Sampling and Estimation for (Sparse) Exchangeable Graphs". In: (2016). arXiv: 1611.00843.

A Overview of Proofs

The appendix is devoted to proving the theoretical results of the paper. These results are obtained subject to the assumption that the data is collected by p-sampling. This assumption is natural in the sense that it provides a reasonable middle ground between a realistic data collection assumption—p-sampling can result in complex models capturing many important graph phenomena [3, 5, 1]—and mathematical tractability—we are able to establish precise guarantees.

The appendix is organized as follows. We begin by recalling the connection between p-sampling and *graphex processes* in Appendix B.1; this affords a useful explicit representation of the data generating process. We recall the method of exchangeable pairs in Appendix B.2. Next, in Appendix B.3, we collect the necessary notation and definitions. Empirical risk convergence results for p-sampling are then proved in Appendix C and results for the random-walk in Appendix D. Finally, convergence results for the global parameters are established in Appendix E.

B Preliminaries

B.1 Graphex processes

Recall the setup for the theoretical results: we consider a very large population network P_t with t edges, and we study the graph-valued stochastic process $(G_n^t)_{n\in[0,\sqrt{t})}$ given taking each G_n to be an $n/\sqrt{t}$-sample from P_t and requiring these samples to cohere in the obvious way. We idealize the population size as infinite by taking the limit $n \to \infty$. The limiting stochastic process $(G_n)_{n\in\mathbb{R}_+}$ is well defined, and is called a *graphex process* [2].

Graphex processes have a convenient explicit representation in terms of (generalized) *graphons* [5, 1, 3].

Definition B.1. A *graphon* is an integrable function $W : \mathbb{R}_+^2 \to [0, 1]$.

Remark B.2. This notion of graphon is somewhat more restricted than graphons (or graphexes) considered in full generality, but it suffices for our purposes and avoids some technical details.

We now describe the generative model for a graphex process with graphon W. Let $\Pi = \{\eta_i\} = \{(\nu(\eta_i), x(\eta_i))\}_{i\in\mathbb{N}}$ be a Poisson (point) process on $\mathbb{R}_+ \times \mathbb{R}_+$ with intensity $\Lambda \otimes \Lambda$, where Λ is the Lebesgue measure. Each atom of the point process is a candidate vertex of the sampled graph; the $\{\nu_i\}$ are interpreted as (real-valued) labels of the vertices, and the $\{x_i\}$ as latent features that explain the graph structure. Each pair of points (η_i, η_j) with $i \le j$ is connected independently with probability $W(x_i, x_j)$. This produces a graph with infinite sample size. To produce a finite sample of size n, we restrict to the collection of edges $\Gamma_n = \{(\eta_i, \eta_j) \ : \ \eta_i, \eta_j \le n\}$. That is, we report the subgraph induced by restricting to vertices with label less than n, and removing all vertices that do not connect to any edges in the subgraph. This last step is critical; in general there are an infinite number of points of the Poisson process such that $\eta_i < n$, but only a finite number of them will connect to any edge in the induced subgraph.

Modeling G_n as collected by p-sampling is equivalent to positing that G_n is the graph structure of Γ_n generated by some graphon W.

B.2 Technical Background: Exchangeable Pairs

We will need to bound the deviation of the normalized degree of a vertice to its expectation. In this goal we here introduce the machinary used to derive those bounds.

A pair of real random variables (X, X') is said to be exchangeable if

$$(X, X') \overset{d}{=} (X', X).$$

Let $f : \mathbb{R} \to \mathbb{R}$ and $F : \mathbb{R}^2 \to \mathbb{R}$ be measurable function such that:

$$\mathbb{E}(F(X, X')|X) \overset{a.s}{=} f(X), \text{ and } \mathrm{F}(\mathrm{X}, \mathrm{X}') = -\mathrm{F}(\mathrm{X}', \mathrm{X}).$$

Let

$$v(X) \triangleq \frac{1}{2}\mathbb{E}\Big(\big(f(X) - f(X')\big)F(X, X')\Big|X\Big),$$

and suppose that $|v(X)| \overset{a.s}{\leq} C$ for a certain $C \in \mathbb{R}$. Then

$$\forall x > 0,\ P(|f(X) - \mathbb{E}(f(X))| \geq x) \leq 2e^{-\frac{x^2}{2C}},$$

and $\forall p > 1$ and $x > 0$ we have the following:

$$P(|f(X) - \mathbb{E}(f(X))| > x) \leq \frac{(2p-1)^p \|v(X)\|_p^p}{x^p}.$$

To see this in more detail see [4].

B.3 Notation

For convenient reference, we collect and explain important notation.

- $\Pi = \{\eta_i = (\nu(\eta_i), x(\eta_i))\}$ is the latent Poisson process that defines the graphex process in Appendix B.1. The labels are ν and the latent variables are x.
- $\Pi_n \triangleq \Pi \cap [0, n] \times \mathbb{R}^+$ is the restriction of the Poisson process to atoms with labels in $[0, n]$.
- $\Gamma_n \subset \mathbb{R}_+^2$ is the (random) edge set of the graphex process at size n.
- $V(\Gamma_n) \subset \mathbb{R}_+$ is the set of vertices of Γ_n.
- $\bar{\Gamma}_n = \{(\eta_i, \eta_j)\ :\ \eta_i, \eta_j \in V(\Gamma_n) \text{ and } (\eta_i, \eta_j) \notin \Gamma_n\}$ is all pairs of points in Γ_n that are not connected by an edge.
- The number of edges in the graph is $E_n = |\Gamma_n|$
- For all k, the set of paths of length k in Γ_n is

$$\mathcal{P}_k(\Gamma_n) \triangleq \{(\eta_i)_{i \leq k+1} \in V(\Gamma_n)^{k+1}\ :\ (\eta_i, \eta_{i+1}) \in \Gamma_n\ \forall i \leq k\}.$$

- The degree of ν in Γ_n is $d_n(\eta)$.
- For mathematical convenience, we attach the embedding vectors to the points of the latent Poisson process. The collection of all possible embeddings is:

$$\Omega_\theta^\Pi \triangleq \{(\lambda_\eta, \gamma)_{\eta \in \Pi}\ :\ \lambda_\eta \in \Omega_\theta\ \forall \eta \in \Pi \text{ and } \gamma \in \Omega_\gamma\}$$

  For all $\bar{\theta} \in \Omega_\theta^\Pi$ we will note $\lambda(\bar{\theta})$ its projection on Ω_λ^Π and respectively $\gamma(\bar{\theta})$ its projection on Ω_γ.
- Asymptotically, the number of edges of a graphex process scales as n^2 [1]. Let $\mathcal{E} \in \mathbb{R}_+$ be the proportionality constant

$$\mathcal{E} \triangleq \lim_{n \to \infty} \frac{E_n}{n^2}.$$

- To build the graph from the point of process Π_n we need to introduce a process of independent uniform variables: $\mathbb{U}_\Pi \triangleq (U_{\eta_i, \eta_j})_{\eta_i, \eta_j \in \Pi}$ be such that $\mathbb{U}_\Pi | \Pi$ is an independent process s.t

$$\forall \eta_1, \eta_2 \in \Pi,\ U_{\eta_1, \eta_2} \sim \text{unif}(0, 1).$$

- Let us introduce the following concepts and notations to build marking on the point process: Let $m(\cdot, \cdot)$ be a distributional kernel on $\mathbb{R}_+ \times \Omega_\theta$, then we generate the marks according to a distribution $\mathcal{Q}_\theta^\Pi$ on Ω_θ^Π, conditional on Π, such that if $\bar{\theta} | \Pi \sim \mathcal{Q}_\theta^\Pi$ then:
  - $(\bar{\theta}_\eta)_{\eta \in \Pi}$ is an independent process
  - $\forall \eta \in \Pi,\ \bar{\theta}_\eta | \Pi \sim m(x(\eta), \cdot)$.
- For all n and $\eta \in \Pi$ we can denote the neighbours of η in Γ_n

$$\mathcal{N}_n(\eta) \triangleq \{\eta' \text{ s.t } (\eta, \eta') \in \mathcal{P}_1(\Gamma_n)\}$$

- For all n we can denote $\bar{\Pi}_n(\theta) \triangleq (\Pi_n, \mathbb{U}|_{\Pi_n}, \theta|_n)$ the augmented object that will cary information not only about the graph structure but also about the markings θ.

C Results for p-sampling

We begin by establishing the result for p-sampling, with the negative examples chosen according to the induced subgraph. This is the simplest case, and is useful for the introduction of ideas and notation. We consider more general approaches to negative sampling to the next section, where it is treated in tandem with random walk sampling. The same arguments can be used to extend p-sampling to allow for, e.g., unigram negative sampling used in our experiments.

For all $\bar{\theta} \in \Omega_\theta^\Pi$ let $\Gamma_n(\bar{\theta})$ be the graph Γ_n where the vertices are annotated with their embeddings from $\bar{\theta}$. We write the empirical risk as $\hat{R}_k(\Gamma_n(\bar{\theta}))$.

Theorem C.1. *Let $\bar{\theta}$ a random variable taking value in Ω_θ^Π s.t $\bar{\theta}\,|\Pi \sim \mathcal{Q}_\theta^\Pi$, for a certain kernel m, then there is some constant $c_m^{\mathrm{ps}} \in \mathbb{R}_+$ such that if $\|\mathcal{L}\|_\infty < \infty$ then*

$$\hat{R}(\Gamma_n(\bar{\theta})) \to c_m^{\mathrm{ps}}$$

both a.s. and in L_1, as $n \to \infty$.

Moreover there is some constant $c_*^{\mathrm{ps}} \in \mathbb{R}_+$ such that

$$\min_\theta \hat{R}(\Gamma_n(\theta)) \to c_*^{\mathrm{ps}}$$

both a.s.and in L_1, as $n \to \infty$.

Proof. We will first prove the first statement. Let $\bar{\theta}|\Pi \sim \mathcal{Q}_\theta^\Pi$, we will note $\Gamma(\bar{\theta})$ the edge set of $\bar{\Pi}(\bar{\theta})$ and denote note $\Gamma^n(\bar{\theta})$ the partially labeled graph obtained from $\Gamma(\bar{\theta})$ by forgetting all labels in $[0, n)$ (but keeping larger labels and the embeddings θ). Let $\mathcal{F}_n(\bar{\theta})$ be the σ-field generated by $\Gamma^n(\bar{\theta})$. A key observation is

$$\hat{R}_k(\Gamma_n(\bar{\theta})) = \mathbb{E}[L(\Gamma_k, \Gamma(\bar{\theta})|_k) \mid \mathcal{F}_n(\bar{\theta})].$$

The reason is that choosing a graph by k/n-sampling is equivalent uniformly relabeling the vertices in $[0, n)$ and restricting to labels less than k; averaging over this random relabeling operation is precisely the expectation on the righthand side.

By the reverse martingale convergence theorem we get that:

$$\hat{R}_k(\Gamma_n(\bar{\theta})) \xrightarrow{a.s, L_1} \mathbb{E}[L(\Gamma_k, \Gamma(\bar{\theta})|_k) \mid \mathcal{F}_\infty(\bar{\theta})],$$

but as $\mathcal{F}_\infty(\bar{\theta})$ is a trivial sigma-algebra we get the desired result.

We will now prove the second statement. For this let us define two new notations: (i) Let Γ^n be the partially labeled graph obtained from Γ by forgetting all labels in $[0, n)$ and let $\mathcal{F}_n$ be the σ-field generated by Γ^n. And (ii) we can also for all $m \in \mathbb{N}$ write the set of embeddings on the graph Γ^m:

$$\Omega_\theta^{\Gamma^m} \triangleq \{(\lambda_{\mathcal{V},\gamma})_{\mathcal{V} \in \Gamma^m} : \forall \mathcal{V} \in V(\Gamma^m)\ \lambda_\mathcal{V} \in \Omega_\lambda, \gamma \in \Omega_\gamma\}.$$

We are now ready to state the proof. Let $m \leq n$, and observe that:

$$\mathbb{E}[\min_{\theta \in \Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n(\theta)) \mid \mathcal{F}_m] \leq \min_{\theta \in \Omega_\theta^{\Gamma^m}} \mathbb{E}[L(\Gamma_k, \Gamma(\theta)|_k) \mid \mathcal{F}_m] \tag{9}$$

$$= \min_{\theta \in \Omega_\theta^{\Gamma^m}} \hat{R}_k(\Gamma_m(\theta)). \tag{10}$$

Thus, $(\min_{\theta \in \Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n(\theta)))_{n \in \mathbb{R}_+}$ is a supermartingale with respect to the filtration $(\mathcal{F}_n)_{n \in \mathbb{R}_+}$. Moreover, by assumption, the loss is bounded and thus so also is the empirical risk. Supermartingale convergence then establishes that $\min_{\theta \in \Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n(\theta))$ converges almost surely and in L_1 to some random variable that is measureable with respect to $\mathcal{F}_\infty$. The proof is completed by the fact that $\mathcal{F}_\infty$ is trivial. □

D Random-walk sampling

In this section we establish the convergence of the relational empirical risk defined by the random walk. The argument proceeds as follows: We first recast the subsampling algorithm as a random

probability measure, measurable with respect to the dataset graph Γ_n. Producing a graph according to the sampling algorithm is the same as drawing a graph according to the random measure. Establishing that the relational empirical risk converges then amounts to establishing that expectations with respect to this random measure converge; this is the content of Lemma D.8. To establish this result, we show in Lemma D.6 that sampling from the random-walk random measure is asymptotically equivalent to a simpler sampling procedure that depends only on the properties of the graphex process and not on the details of the dataset. We allow for very general negative sampling distributions in this result; we show that how to specialize to the important case of (a power of) the unigram distribution in Lemma D.7.

D.1 Random-walk Notation

We begin with a formal description of the subsampling procedure that defines the relational empirical risk. We will work with random subset of the Poisson process Π; these translate to random subgraphs of Γ in the obvious way. Namely, if the sampler selects $\eta_i = (\nu_i, x_i)$ in the Poisson process, then it selects η_i in Γ.

Sampling follows a two stage procedure: we first choose a random walk, and then augment this random walk with additional vertices—intuitively, this is the negative-sampling step. The following introduces much of the notation we require for this section.

Definition D.1 (Random-walk sampler)**.** Let μ_n be a (random) probability measure over $V(\Gamma_n)$. Let $H = (\eta_i)_{i \leq M} = (\nu(\eta_i), \lambda(\eta_i))_{i \leq M}$ be a sequence of vertices sampled according to:

1. (random-walk) $\eta_1 \sim \frac{d_n(\eta_1)}{2E_n}$ and let $\eta_i | \eta_{i-1} \sim \mathrm{unif}(\mathcal{N}_{\mathrm{n}}(\eta_{\mathrm{i}-1}))$ for $i \in (2, \ldots, r+1)$.
2. (augmentation) $\eta_{r+2:M}$ be a sequence of additional vertices sampled from μ_n independently from each other and also from $(\eta_1, \ldots, \eta_{r+1})$.

Let G_H be the vertex induced subgraph of Γ_n. Let $P_n = \mathbb{P}(\Gamma_H \in \cdot \mid \bar{\Pi}_n(\bar{\theta}))$ be the random probability over subgraphs induced by this sampling scheme. Finally, let $G_H(\bar{\theta})$ denote the same subgraph augmented with the embeddings.

With this notation in hand, We rewrite the loss function and the risk in a mathematically convenient form

Definition D.2 (Loss and risk)**.** The loss on a subsample is

$$L(G_H, G_H(\bar{\theta})) \in [0, 1].$$

The empirical risk is

$$\mathbb{E}_{P_n}[L(G_H(\lambda), G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})].$$

Remark D.3. Note that the subgraphs produced by the sampling algorithm explicitly include all edges and non-edges of the graph. However, the loss may (and generally will) depend on only a subset of the pairs. In this fashion, we allow for the practically necessary division between negative and positive examples. Skipgram augmentation can be handled with the same strategy.

We impose a technical condition on the distribution that the additional vertices are drawn from. Intuitively, the condition is that the distribution is not too sensitive to details of the dataset in the large data limit.

Definition D.4 (Augmentation distribution)**.** We say μ_n is an *asymptotically exchangeable augmentation distribution* if is there is a μ such that

- There is a deterministic function f s.t $\mu(\eta) = f(\theta(\eta))$
- $\|\mu_n(\cdot) - \frac{\mu(\cdot)\mathbb{I}(\cdot \in \Gamma_n)}{n\ Z_n}\|_{TV} \xrightarrow{P} 0$, where $Z_n \triangleq \frac{1}{n}\sum_{\eta \in \Pi_n} \mu(\eta)$.

We will later prove that the unigram distributions respect that.

D.2 Technical lemmas

We begin with some technical inequalities controlling sums over the latent Poisson process.

To bound the remaining quantities we will first prove that the following is true:

Lemma D.5. *Let $(U_{x(\eta)})_{\eta\in\Pi}$ is such that $(U_{x(\eta)})_{\eta\in\Pi}|\Pi$ be distributed as a process of independent uniforms in $[0,1]$ and let*

$$\forall y \in \mathbb{R}_+, \quad f_n(y,\Pi) \triangleq \sum_{\eta\in\Pi_n} \mathbb{I}(U_{x(\eta)} \leq W(y,x)).$$

Then the following hold:

1. *$\forall y \in \mathbb{R}_+$ such that $W(y,\cdot) \geq n^{-1+\frac{\epsilon}{4}}$, there are $p, K > 0$ such that $\forall \beta > 0$,*

$$\mathbb{P}(\Big|\frac{f_n(y,\Pi)}{nW(y,\cdot)} - 1\Big| \geq \beta) \leq \frac{K}{n^3\beta^p}.$$

2. *$\forall p > 0,\ \exists K_p$ such that $\forall \beta > 0$*

$$\mathbb{P}(\Big|\frac{f_n(y,\Pi))}{n} - W(y,\cdot)\Big| \geq \beta) \leq \frac{K_p}{n^p\beta^{2p}}$$

and

$$\mathbb{P}(\Big|\frac{E_n}{n^2\mathcal{E}} - 1\Big| \geq \beta) \leq \frac{K_p}{n^p\beta^{2p}}.$$

3. *$\exists K \in \mathbb{R}_+$ such that $\forall y \in \mathbb{R}_+$, such that $W(y,\cdot) \leq n^{-1+\frac{\epsilon}{4}}$ then $\mathbb{P}(f_n(\Pi, y) \geq n^{\frac{\epsilon}{2}}) \leq \frac{K}{n^3}$.*

Proof. We will first write the proof of the first statement, which is harder. We then highlight the differences in the other cases. We use the Stein exchangeable pair method, previously presented.

Let $F : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\forall x, y\ F(x,y) = [x - y].$$

Let $\bar{J} \sim \text{unif}(\{0, \text{n} - 1\})$ and let

$$\Pi' = T_{[\bar{J},\bar{J}+1],[n,n+1]} \cdot \Pi_\nu \times \Pi_x,$$

where $T_{[\bar{J},\bar{J}+1],[n,n+1]}$ is the permutation of $[\bar{J}, \bar{J}+1]$ and $[n, n+1]$ and

$$T_{[\bar{J},\bar{J}+1],[n,n+1]} \cdot \Pi_\nu \times \Pi_x \triangleq \{(T_{[\bar{J},\bar{J}+1],[n,n+1]}(\nu), x),\ \forall(\nu,x) \in \Pi.\}$$

Then we can check the following:

- As $\Pi \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+ = \Pi' \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+$ we obtain that

$$\begin{aligned}
&\mathbb{E}(\frac{f_n(y,\Pi)}{W(y,\cdot)} - \frac{f_n(y,\Pi')}{W(y,\cdot)}|\Pi_n) \\
&\overset{(a)}{=} \frac{1}{nW(y,\cdot)}[\sum_{j=0}^{n-1} \sum_{\Pi_{j+1}\setminus\Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x)) - \mathbb{E}(\mathbb{I}(U_{x(\eta)} \leq W(y,x)))] \\
&\overset{(b)}{=} \frac{f_n(y,\Pi)}{nW(y,\cdot)} - 1
\end{aligned}$$

  where (a) is obtained by complete independence of Π and where to get (b) we use the fact that as mentionned in [5] we know that

$$\forall j, \sum_{(\nu,x)\in\Pi_{j+1}\setminus\Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x)) \sim \text{POI(W(y,}\cdot\text{))}$$

- Moreover, we can very similarly see that:

$$\begin{aligned}
&\Big\|\frac{1}{2n}\mathbb{E}([\frac{f_n(y,\Pi)}{W(y,\cdot)}-\frac{f_n(y,\Pi')}{W(y,\cdot)}]^2\big|\Pi_n)\Big\|_p \\
&\leq \frac{1}{n^2W(y,\cdot)^2}\Big\|\sum_{j=0}^{n-1}\big[[\sum_{(\nu,x)\in\Pi_{j+1}\setminus\Pi_j}\mathbb{I}(U_{x(\eta)}\leq W(y,x))]^2+2W(y,\cdot)\big]\Big\|_p \\
&\leq \frac{1}{n^2W(y,\cdot)^2}\sum_{j=0}^{n-1}\Big\|[\sum_{(\nu,x)\in\Pi_{j+1}\setminus\Pi_j}\mathbb{I}(U_{x(\eta)}\leq W(y,x))]^2\Big\|_p+2W(y,\cdot) \\
&\leq \frac{C}{nW(y,\cdot)},
\end{aligned}$$

  where C is a constant that does not depend on n or y.

Therefore using the exchangeable pair method presented earlier and setting $p \geq \frac{12}{\epsilon}$ for all y s.t $W(y,\cdot) \geq n^{\frac{\epsilon}{4}-1}$ we get that there is K,p s.t for all $\epsilon > 0$

$$P(|\frac{\sum_{(\nu,x)\in\Pi_n}\mathbb{I}(U_{x(\eta)}\leq W(y,x))}{W(y,\cdot)}-1|\geq\beta)\leq\frac{K}{n^3\beta^p},$$

QED.

For the second statement, instead of $\frac{f_n(y,\Pi)}{W(y,\cdot)}$ we are interested in simply $f_n(y,\Pi)$ which is easier to deal. Indeed, using the same exchangeable pair (Π,Π') we get that:

- As $\Pi\cap[0,n]\setminus[\bar{j},\bar{j}+1]\times\mathbb{R}^+=\Pi'\cap[0,n]\setminus[\bar{j},\bar{j}+1]\times\mathbb{R}^+$ we obtain that

$$\begin{aligned}
&\mathbb{E}(f_n(y,\Pi)-f_n(y,\Pi')\big|\Pi_n) \\
&=\frac{1}{n}f_n(y,\Pi)-W(y,\cdot).
\end{aligned}$$

- Moreover we can very similarly see that:

$$\begin{aligned}
&\Big\|\frac{1}{2n}\mathbb{E}([f_n(y,\Pi)-f_n(y,\Pi')]^2\big|\Pi_n)\Big\|_p \\
&\leq \frac{1}{n^2}\sum_{j=0}^{n-1}\Big\|[\sum_{(\nu,x)\in\Pi_{j+1}\setminus\Pi_j}\mathbb{I}(U_{x(\eta)}\leq W(y,x))]^2\Big\|_p+2W(y,\cdot) \\
&\leq \frac{C}{n},
\end{aligned}$$

  where C is a constant that does not depend on n or y. Therefore we get the desired result QED.

  A very similar roadmap can be followed for N_e^n.

  The last statement is a simple consequence of this. Indeed, $\forall y\in\mathbb{R}$

$$P(W(y)\leq n^{-1+\frac{\epsilon}{4}},f_n(\Pi,y)\geq n^{\frac{\epsilon}{2}})\leq P(|\frac{f_n(\Pi,y)}{n}-W(y,\cdot)|\geq n^{-\frac{\epsilon}{4}})\leq\frac{K_{\frac{3}{1+\frac{\epsilon}{4}}}}{n^3}.$$

□

With this in hand, we establish the asymptotic equivalence of random-walk sampling and a sampling scheme that does not depend on the details of the dataset. This is the key component of the proof. Recall the notation introduced in Appendix D.1.

Lemma D.6. *Suppose that there is $\epsilon \in (0,1)$ such that the graphon W verifies*

$$W(x,\cdot) = O(x^{-1-\epsilon}).$$

Suppose further that the augmented sampling distributions $(\mu_n)_n$ satisfy the conditions of Definition D.4. Then, writing

$$P_n(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu_n(\eta_l)}{2N_e^n \prod_{i=2}^{r} d_n(\eta_i)}$$

and

$$\tilde{P}_n(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu(\eta_l)}{2n^M \mathcal{E} \prod_{i=2}^{r} W(x(\eta_i),\cdot)},$$

it holds that

$$\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| = o_p(1).$$

Proof. We can first see by the triangle inequality that if we write the following two measures:

$$P_n^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu(\eta_l)}{2N_e^n n^{M-(r+1)} \prod_{i=2}^{r} d_n(\eta_i)}$$

and

$$\tilde{P}_n^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{i=2}^{r} \mathbb{I}(W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}}) \prod_{l=r+2}^{M} \mu(\eta_l)}{2n^M \mathcal{E} \prod_{i=2}^{r} W(x(\eta_i),\cdot)}$$

Then $\forall \beta > 0$:

$$\begin{aligned}
&P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| > \beta \Big) \\
&\leq P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| > \frac{\beta}{3} \Big) \\
&+ P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| > \frac{\beta}{3} \Big) \\
&+ P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| > \frac{\beta}{3} \Big),
\end{aligned}$$

therefore proving that those last terms converge to zero for any $\beta > 0$ is sufficient.

First we will prove that

$$\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| = o_p(1).$$

Indeed, noting that,

$$P_{n,i}^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{r+1+i} \mu(\eta_l) \prod_{r+2+i}^{M} \mu_n(\eta_l)}{2N_e^n n^i \prod_{i=2}^{r} d_n(\eta_i)},$$

it holds $\forall \beta > 0$ that

$$\begin{aligned}
&P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) \right| > \beta \Big) \\
&\overset{(a)}{\leq} \sum_{i=1}^{M} P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \left| \mathbb{E}_{P_{n,i}^*}\big(L(G_H, G_H(\bar{\theta})\big) - \mathbb{E}_{P_{n,i-1}^*}\big(L(G_H, G_H(\bar{\theta})\big) \right| > \frac{\beta}{M} \Big) \\
&\leq M P\big(\|\mu_n - \frac{\mu}{nZ_\mu}\|_{TV} > \frac{\beta}{\|L\|_\infty} \big).
\end{aligned}$$

where (a) using telescopic sum. Therefore we have proven that the first element of the sum goes to 0.

Now we will prove that

$$\sup_{\bar{\theta}\in\Omega_{\theta}^{\Pi}} \mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})\Big|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big)\Big| = o_p(1).$$

For this we will want to approximate $\frac{n}{d_n(V_{u_i})}$ by $\frac{1}{W(u_i,\cdot)}$. However for this we need a good bound on $P(|\frac{d_n(V_{u_i})}{nW(u_i,\cdot)} - 1| \geq \epsilon)$. But this is possible only if $W(u_i,\cdot)$ is not too small.

But note that for all vertices $\eta \in \Pi_n$ if a path H passes through η at the i-th coordinate, for $i \geq 2$, then it means that there is only $d_n(\nu(\eta))$ possibilities for the $i-1$th vertex of the path. Therefore if $d_n(\nu(\eta))$ is small the probability that our random-walk passes through v, and is not the origin vertex, is asymptotically negligible.

Indeed for all $\eta \in \Pi_n$ such that $d_n(\nu(\eta)) \leq n^{\frac{\epsilon}{2}}$ $\forall k \geq 2$

$$P(\eta_i = \eta|\bar{\Pi}_n(\bar{\theta})) \leq \sum_{\eta'\in\Pi_n\cap\mathcal{N}_n(\eta)} P(\eta_{i-1}=\eta', \eta_i=\eta|\bar{\Pi}_n(\bar{\theta})) \overset{(*)}{\leq} \frac{n^{\frac{\epsilon}{2}}}{2N_n^e},$$

where to get (*) we used the stationary property of the RW.

Therefore we have:

$$P(\min_{k\geq 2} d_n(\eta_k) \leq n^{-\frac{\epsilon}{2}}|\bar{\Pi}_n(\bar{\theta})) \leq \frac{rn^{\frac{\epsilon}{2}}\big|\{\eta\in\Pi_n, \text{ s.t. } 0 < d_n(\eta) \leq n^{\frac{\epsilon}{2}}\}\big|}{2N_n^e} \xrightarrow{P} 0,$$

But we have that $\forall(\eta_i)_{i\leq r+1}$ s.t. $\forall i, W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}}$,

$$\begin{aligned}
&\Big|\frac{1}{2N_e^n\prod_{i=2}^r d_n(\eta_i)} - \frac{1}{2n^{r+1}\mathcal{E}\prod_{i=2}^r W(x(\eta_i),\cdot)}\Big| \\
&\overset{(a)}{\leq} \sum_{i=2}^r \frac{1}{2N_e^n n^{i-1}\prod_{l=2}^{r-i} d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\Big|\frac{1}{d_n(\eta_{r-i+1})} - \frac{1}{nW(x(\eta_{r-i+1}),\cdot)}\Big| \\
&\quad + \frac{1}{n^{r-1}\prod_{l=2}^r W(x(\eta_l),\cdot)}\Big|\frac{1}{2N_e^n} - \frac{1}{2n^2\mathcal{E}}\Big| \\
&\leq \sum_{i=2}^r \frac{1}{2N_e^n n^{i-1}\prod_{l=2}^{r-i+1} d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\Big|1 - \frac{d_n(\eta_{r-i+1})}{nW(x(\eta_{r-i+1}),\cdot)}\Big| + \frac{1}{2n^{r-1}N_e^n\prod_{l=2}^r W(x(\eta_l),\cdot)}\Big|1 - \frac{N_e^n}{n^2\mathcal{E}}\Big|,
\end{aligned}$$

where (a) comes from a simple telescopic sum re-writing.

Therefore if

$$\max_i \Big|1 - \frac{d_n(\nu_i)}{nW(y_i,\cdot)}\Big|, \Big|1 - \frac{N_e^n}{n^2\mathcal{E}}\Big| \leq \beta$$

then

$$\begin{aligned}
&\Big|\frac{1}{2N_e^n\prod_{i=2}^r d_n(\eta_i)} - \frac{1}{2n^{r+1}\mathcal{E}\prod_{i=1}^r W(x(\eta_i),\cdot)}\Big| \\
&\leq \beta\big[\sum_{i=2}^r \frac{1}{2N_e^n n^{i-1}\prod_{l=2}^{r-i+1} d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)} + \frac{1}{2n^{r-1}N_e^n\prod_{l=2}^r W(x(\eta_l),\cdot)}\big]
\end{aligned}$$

Now note that for all i, and $\lambda' \in \Omega$

$$\begin{aligned}
&\sum_{\eta_{1:r+1}\in\mathcal{P}_r(\Pi_n)} \frac{\prod_{i=2}^r \mathbb{I}(W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}})}{2N_e^n n^{i-1}\prod_{l=2}^{r-i+1} d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\mathbb{E}\big(L(G_H, G_H(\bar{\theta})|\eta_{r+2:M_n}, \Pi_n\big) \\
&\overset{(a)}{\leq} \sum_{\eta_{1:r}\in\mathcal{P}_{r-1}(\Pi_n)} d_n(\eta_r)\frac{\prod_{i=2}^r \mathbb{I}(W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}})}{2N_e^n n^{i-1}\prod_{l=2}^{r-i+1} d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\mathbb{E}\big(L(G_H, G_H(\bar{\theta})|\eta_{r+2:M_n}, \Pi_n\big) \\
&\leq \|L\|_\infty \max_{y\in N_v^n(\Pi) \text{ s.t. } W(y,\cdot)\geq n^{-1+\frac{\epsilon}{4}}} \frac{d_n(y)}{nW(y,\cdot)} \sum_{\eta_{1:r}\in\mathcal{P}_{r-1}(\Pi_n)} \frac{\prod_{i=2}^r \mathbb{I}(W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}})}{2N_e^n n^{i-1}\prod_{l=2}^{r-i+1} d_n(\eta_l)\prod_{l=r-i+2}^{r-1} W(x(\eta_l),\cdot)}
\end{aligned}$$

where (a) is a simple consequence form the fact that:

$$\mathrm{card}\{\eta \in \eta(\Pi_\mathrm{n}, \mathrm{r}) \,\mathrm{s.t.}\, \eta|_{1:\mathrm{r}} = (\nu_\mathrm{i}, \mathrm{y_i})_{1:\mathrm{r}}\} = \mathrm{d_n}(\nu_\mathrm{r})\mathrm{card}\{\eta \in \eta(\Pi_\mathrm{n}, \mathrm{r}-1) \,\mathrm{s.t.}\, \eta|_{1:\mathrm{r}-1} = (\nu_\mathrm{i}, \mathrm{y_i})_{1:\mathrm{r}-1}\}.$$

Therefore, by induction, we can get that for all i

$$\sum_{\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)} \prod_{i=2}^{r} \mathbb{I}(W(x(\eta_i), \cdot) \geq n^{-1+\frac{\epsilon}{4}}) \frac{\mathbb{E}\big(L(G_H, G_H(\bar{\theta})|\eta_{r+2:M}, \Pi_n\big)}{N_e^n n^{i-1} \prod_{l=2}^{r-i+1} d_n(\eta_l) \prod_{l=r-i+2}^{r} W(x(\eta_l), \cdot)}$$

$$\leq r\|L\|_\infty \max_{y \in N_v^n(y) \;\; \mathrm{s.t.}\; W(y,\cdot) \geq n^{-1+\frac{\epsilon}{4}}} |\frac{d_n(y)}{nW(y,\cdot)} - 1| + \|L\|_\infty.$$

Therefore if we note

$$A_n(\beta) \triangleq \{ \max_{y \in N_v^n(y) \;\; \mathrm{s.t.}\; W(y,\cdot) \geq n^{-1+\frac{\epsilon}{4}}} |\frac{d_n(y)}{nW(y,\cdot)} - 1| \leq \beta, |\frac{N_e^n}{n^2\mathcal{E}} - 1| \leq \beta\}$$

Then we can see the following:

- On $A_n(\beta)$ we will have that as $\eta_{1:r+1} \perp\!\!\!\perp \eta_{r+2:M}$ using the result that we previously got we have that:

$$\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big|\mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big)\Big| \leq (r+1)^2\|L\|_\infty\beta$$

- And in addition we know that there is $K_1, K_2 < \infty$ s.t

$$P(A_n(\beta)^c) \leq P(|\frac{N_e^n}{n^2\mathcal{E}} - 1| \geq \beta) + \mathbb{E}\big(\sum_{\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)} \mathbb{I}(|\frac{d_n(y)}{nW(y,\cdot)} - 1| \geq \beta)\big)$$

$$\overset{(a)}{\leq} P(|\frac{N_e^n}{n^2\mathcal{E}} - 1| \geq \beta) + n\int_{\mathbb{R}^+} \mathbb{I}(W(x,\cdot) \geq n^{-1+\frac{\epsilon}{4}}) P\big(|\frac{f_n(x,\Pi)}{nW(x,\cdot)} - 1| \geq \beta\big)\big)dx$$

$$\overset{(b)}{\leq} \frac{K_1}{n\beta} + \frac{K_2}{\beta^p n^2}\int_{\mathbb{R}^+} \mathbb{I}(W(x,\cdot) \geq n^{-1+\frac{\epsilon}{4}})dx$$

$$\leq \frac{K_1}{n\beta} + \frac{K_2}{\beta^p n^2} n^{1-\frac{3\epsilon}{2+2\epsilon}} \to 0,$$

where (a) comes from Slivnyak–Mecke theorem and (b) from our previous lemma Lemma D.5

Thus, we have successfully proven that

$$\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big|\mathbb{E}_{P_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big)\Big| = o_p(1)$$

QED

Now we are going to prove the last part, i.e.

$$\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big|\mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big)\Big| = o_p(1)$$

For this we can note that that for all $i \geq 2$

$$\|\frac{1}{n^{r+1}} \sup_{\lambda' \in \Omega_\theta^\Pi} \sum_{\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)} \frac{\mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}})}{2n^{r+1}\mathcal{E}\prod_{i=2}^{r} W(x(\eta_i),\cdot)} \mathbb{E}\big(L(G_H, G_H(\bar{\theta})|\bar{\Pi}_n(\bar{\theta}), \eta_{r+2,M}\big)\|_{L_1}$$

$$\overset{(a)}{\leq} \|L\|_\infty \int_{\mathbb{R}^{r+1}} \mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}}) \frac{\prod_{j=1}^{r} W(x_j, x_{j+1})}{\prod_{j=2}^{r} W(x_j,\cdot)} dx_{1:r+1}$$

$$\overset{(b)}{\leq} \|L\|_\infty \int_{\mathbb{R}^{i}} \mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}}) \frac{\prod_{j=1}^{i-1} W(x_j, x_{j+1})}{\prod_{j=2}^{i-1} W(x_j,\cdot)} dx_{1:i}$$

$$\overset{(c)}{\leq} \|L\|_\infty \int_{\mathbb{R}} W(x(\eta_i),\cdot)\mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}})dx_i \xrightarrow{n\to\infty} 0,$$

where to get (a) we used both the fact that L was bounded and the independence of the uniforms; to get (b) we integrated coordinates $r+1$ to $i+1$ and used the following definition:

$$\forall x \int W(x', x)dx' = W(x, \cdot).$$

We similarly got (c) where instead we integrated the coordinates from 1 to $i-1$.

Therefore we have successfully proven that

$$\sup_{\bar{\theta} \in \Omega_{\bar{\theta}}^{\Pi}} \left| \mathbb{E}_{\tilde{P}_n^*} \big(L(G_H, G_H(\bar{\theta}) | \bar{\Pi}_n(\bar{\theta}) \big) - \mathbb{E}_{\tilde{P}_n} \big(L(G_H, G_H(\bar{\theta}) | \bar{\Pi}_n(\bar{\theta}) \big) \right| = o_p(1)$$

Hence we have proven the desired results □

We now turn to the question of which augmentation distributions will satisfy the conditions of the previous result. We show that the conditions hold for any distribution defined by a differentiable function of the unigram distribution; in particular, this covers the unigram distribution to the power of $3/4$ that is used to define unigram negative sampling.

Lemma D.7. *Let $\eta_{1:r+1}$ be sampled by a random walk on G_n, and let the random-walk unigram distribution be defined by*

$$\mathrm{Ug}_{\Gamma_n}(\eta) = \mathbb{P}(\exists i \leq r+1, \text{ s.t. } \tilde{\eta}_i = \eta \mid \bar{\Pi}_n(\bar{\lambda})).$$

Suppose that μ_n is defined by

$$\mu_n(\eta) \propto \mathrm{Ug}_{\Gamma_n}(\eta)^{\alpha},$$

for a certain $\alpha > 0$.Then, defining μ by

$$\mu(\eta) \propto (r+1)^{\alpha-1} \frac{W(x, \cdot)^{\alpha-1}}{\mathcal{E}^{\alpha-1}},$$

it holds that

$$\| \mu_n - \frac{\mu(\cdot) \mathbb{I}(\cdot \in \Pi_n)}{n Z_n} \|_{TV} \xrightarrow{p} 0$$

Proof. We will for simplicity prove the result for $\alpha = 1$, the other cases can be obtained following a similarly, although the computations are more involved.

$$\begin{aligned}
&\sum_{\eta \in \Pi_n} \big| P(\exists i \leq r+1, \text{ s.t. } \tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda})) - \sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda})) \big| \\
&\overset{(a)}{\leq} \sum_{\eta \in \Pi_n} \sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda})) P(\exists j \in [i+1, r+1], \ \eta_j = \eta | \eta_i = \eta, \bar{\Pi}_n(\bar{\lambda})) \xrightarrow{P, (b)} 0,
\end{aligned}$$

where (b) comes from the dominated convergence theorem and (a) comes from the fact that for all η

$$\begin{aligned}
&\Big| \mathbb{E}\big(\mathbb{I}\big(\exists i \leq r+1, \text{ s.t. } \tilde{\eta}_i = \eta \big) - \sum_{i=1}^{r+1} \mathbb{I}(\tilde{\eta}_i = \eta) | \bar{\Pi}_n(\bar{\lambda}) \big) \Big| \\
&\leq \sum_{i=1}^{r+1} \mathbb{E}\big(\mathbb{I}(\tilde{\eta}_i = \eta, \ \exists j \geq i \text{ s.t. } \tilde{\eta}_j = \eta) | \Gamma_n \big)
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sum_{\eta \in \Pi_n} \big| \sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda})) - \frac{(r+1) W(x(\eta), \cdot)}{2n\mathcal{E}} \big| \\
&\overset{(a)}{\leq} \sum_{\eta \in \Pi_n} \big| \frac{(r+1) d_n(\eta)}{2 N_e^n} - \frac{(r+1) W(x(\eta), \cdot)}{2n\mathcal{E}} \big|
\end{aligned}$$

where (a) comes from the stationarity proprieties of the simple random walk.

Therefore, using what we have previously done, we see that:

$$\sum_{\eta\in\Pi_n} \Big|\sum_{i=1}^{r+1} P(\tilde{\eta_i}=\eta|\bar{\Pi}_n(\bar{\lambda})) - \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}\Big| = o_p(1).$$

Finally,

$$\begin{aligned}
&\sum_{\eta\in\Pi_n} \Big|\frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}[1-\frac{1}{\sum_{\eta\in\Pi_n}\frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}}]\Big| \\
&= \sum_{\eta\in\Pi_n} \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}} - 1 \\
&= \sum_{\eta\in\Pi_n} \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}} - P(\exists i \leq r+1, \text{ s.t. } \tilde{\eta_i}=\eta|\Gamma_n) = o_p(1).
\end{aligned}$$

□

D.3 Convergence for random walk sampling

Let $\bar{\theta}$ be a random element of Ω_θ^Π such that $\bar{\theta}|\Pi \sim \mathcal{Q}_\theta^\Pi$ for a certain kernel m. For simplicity of exposition we will write

$$\forall n\ R(G_n,\bar{\theta}) \triangleq \mathbb{E}_{P_n}\big(L(G_H,G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big).$$

Lemma D.8. *There are constants $R_{\bar{\theta}}, R^*_{\bar{\theta}} \in \mathbb{R}_+$ such that*

$$R(G_n,\bar{\theta}) \xrightarrow{P} R_m$$

Moreover

$$\min_{\bar{\theta}\in\Omega_w^\Pi} R(G_n,\bar{\theta}) \xrightarrow{P} R^*.$$

And those constants are respectively $\lim_n \mathbb{E}(R(G_n,\bar{\theta}))$ *and* $\lim_n \mathbb{E}(\min_{\bar{\theta}\in\Omega_w^\Pi} R(G_n,\bar{\theta}))$

Proof. Lemma D.6 states that

- $\mathbb{E}_{P_n}\big(L(G_H,G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H,G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big) = o_p(1).$
- $\min_{\bar{\theta}\in\Omega_\theta^\Pi} \mathbb{E}_{P_n}\big(L(G_H,G_H(\lambda))|\bar{\Pi}_n(\bar{\theta})\big) - \min_{\bar{\theta}\in\Omega_\theta^\Pi} \mathbb{E}_{\tilde{P}_n}\big(L(G_H,G_H(\lambda))|\bar{\Pi}_n(\bar{\theta})\big) = o_p(1).$

To see why this is interesting first note $\forall I=(I_1,\dots,I_M)\in\mathbb{N}^M$ the following quantity

$$X_I(\bar{\theta}) \triangleq \sum_{\eta_{1:M}\in\Pi|_I} \frac{\mathbb{I}(\eta_{1:r+1}\in\mathcal{P}_r(\Pi_n))\prod_{l=r+2}^M \mu(x(\eta_l))}{2\mathcal{E}\prod_{i=2}^r W(x(\eta_i),\cdot)} L(G_H,G_H(\bar{\theta})),$$

where $\Pi|_I \triangleq (\Pi_{I_1+1}\setminus\Pi_{I_1},\dots,\Pi_{I_M+1}\setminus\Pi_{I_M})$.

This allows us to write that

$$\mathbb{E}_{\tilde{P}_n}\big(L(G_H,G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big) = \frac{1}{n^M}\sum_{i_{1:M}\leq n-1} X^{\bar{\theta}}_{i_{1:M}},$$

But $(X_I(\bar{\theta}))_{I\in\mathbb{N}^M}$ is an M-dimensional exchangeable array. Therefore using classical results we see that:

$$\mathbb{E}_{P_n}\big(L(G_H, G_H(\bar{\theta}))|\bar{\Pi}_n(\bar{\theta})\big) \xrightarrow{P} \int_{\mathbb{R}_+^M} \mathcal{R}(x_{1:M}) \frac{\prod_{i=r+2}^{M} \mu(x_i)}{2\mathcal{E}\prod_{i=2}^{r} W(x_i,\cdot)} dx_{1:M},$$

where

$$\mathcal{R}(x_{1:M}) = \mathbb{E}\Big(L(G_{x_{1:M}}, G_{x_{1:M}}(\theta_{x_{1:M}}))\prod_{i=1}^{r}\mathbb{I}(U_i \leq W(x_i, x_{i+1}))\Big),$$

and where $G_{x_{1:M}}$ is the subgraph with vertices having intensities respectively $x_1, \ldots, x_m$,and $\forall i,\ \theta_{x_i} \overset{iid}{\sim} m(x_i,\cdot)$.

Now let write for all n, $\mathbb{F}_n$ the sigma-field of events invariant under joint permutations of the indexes in $[1,n]^M$. Then we can see that $(\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{\prod_{i=0}^{M-1}(n-i)} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta}), \mathbb{F}_n)$ is a reverse supermatingale. Indeed

- $\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{\prod_{i=0}^{M-1}(n-i)} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})$ is $\mathbb{F}_n$ measurable as it is invariant under joint permutations of the indexes in $[1,n]^M$.
- For all $m \geq n$ let $\hat{\theta}_m \in \Omega_\theta^{\Pi}$ such that:

$$\sum_{I\in[|1,m-1|]^M} X_I(\hat{\theta}_m) = \min_{\bar{\theta}\in\Omega_\theta^{\Pi_m}} \sum_{I\in[|1,m-1|]^M} X_I(\bar{\theta})$$

  Then we get

$$\begin{aligned}
&\mathbb{E}\Big(\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})|F_m\Big) \\
&\overset{(a)}{\leq} \mathbb{E}\Big(\frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\hat{\theta}_m)|F_m\Big) \\
&\overset{(b)}{\leq} \min_{\bar{\theta}\in\Omega_\theta^{\Pi_m}} \frac{1}{m^M} \sum_{I\in[|1,m-1|]^M} X_I(\bar{\theta}),
\end{aligned}$$

  where (a) comes from Jensen and (b) comes from a standard argument in exchangeable arrays.

Therefore we have that:

$$\min_{\bar{\theta}\in\Omega_\theta^{\Pi}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta}) - \mathbb{E}(\min_{\bar{\theta}\in\Omega_\theta^{\Pi}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})) \xrightarrow{P} 0.$$

□

E Convergence of global parameters

Suppose that Ω_γ is a compact convex set. Let $\epsilon > 0$ we say that the loss function L is ϵ-strictly convex in γ if for all $\gamma, \gamma' \in \Omega_\gamma,\ \forall \eta \in [0,1]$ and for all $\bar{\theta}_\gamma, \bar{\theta}_{\gamma'}, \bar{\theta}_{\eta\gamma'+(1-\eta)\gamma}$ s.t

$$(i)\ \lambda(\bar{\theta}_\gamma) = \lambda(\bar{\theta}_{\gamma'}) = \lambda(\bar{\theta}_{\eta\gamma'+(1-\eta)\gamma}),$$

and

$$(ii)\ \gamma(\bar{\theta}_\gamma) = \gamma,\ \gamma(\bar{\theta}_{\gamma'}) = \gamma',\ \gamma(\bar{\theta}_{(1-\eta)\gamma+\eta\gamma'}) = (1-\eta)\gamma + \eta\gamma',$$

we have

$$L(G_H, G_H(\bar{\theta}_{\eta\gamma'+(1-\eta)\gamma})) \overset{a.s}{<} \eta L(G_H, G_H(\bar{\theta}_{\gamma'})) + (1-\eta)L(G_H, G_H(\bar{\theta}_\gamma)) - \epsilon.$$

We will moreover say that L is uniformly-continuous if

$$\lim_{\gamma' \to \gamma} \left\| \sup_{\bar{\lambda} \in \Omega_\lambda^\Pi} \left| L(G_H, G_H(\bar{\theta}_{\gamma'})) - L(G_H, G_H(\bar{\theta}_\gamma)) \right| \right\|_{L_1} = 0.$$

For ease of presentations we will write the risk slightly differently i.e: $\forall \bar{\theta} \in \Omega_\theta^\Pi$ s.t $\gamma(\bar{\theta}) = \gamma$ and $\lambda(\bar{\theta}) = \lambda$ we will write the risk $\forall s \in \mathbb{R},\ \mathcal{R}_r(\gamma, \lambda, G_s)$, indifferent of which sampler we are using.

Lemma E.1. *Suppose that there is $\epsilon > 0$ such that L is ϵ-strictly convex and uniformly-continuous in γ, suppose that Ω_γ is a compact convex set. Let $(\hat{\gamma}_n)_n \in \Omega_\gamma^{\mathbb{N}}$ be a sequence of elements in Ω_γ such that*

$$\forall n,\ \min_{\lambda \in \Omega_\lambda^\Pi} \mathcal{R}_r(G_n, \hat{\gamma}_n, \lambda) = \min_{\gamma \in \Omega_\gamma} \min_{\lambda \in \Omega_\lambda^\Pi} \mathcal{R}_r(G_n, \gamma, \lambda).$$

Then there is γ^* s.t

$$\hat{\gamma_n} \xrightarrow{P} \gamma^*,$$

where $\gamma^* = \mathrm{argmin}_\gamma \mathrm{lim}_\mathrm{n}\, \mathbb{E}(\min_{\lambda \in \Omega_\lambda^\Pi} \mathcal{R}_\mathrm{r}(\mathrm{G_n}, \gamma, \lambda))$

Remark E.2. This result is valid for both random-walk and p-sampling.

Proof. Let write $\mathcal{R}_r(\gamma, G_s) \triangleq \min_{\lambda \in \Omega_\theta^\Pi} \mathcal{R}_r(\gamma, \lambda, G_s)$.

Lemma D.8 and Theorem C.1 respectively for the random-walk sampler and p-sampling gives us the following result for all γ,

$$\mathcal{R}_r(\gamma, G_s) - \mathbb{E}(\mathcal{R}_r(\gamma, G_s))) \xrightarrow{P} 0.$$

Let note $(\hat{\gamma}_n)_n \in \Omega_\gamma^{\mathbb{N}}$ be a sequence such that

$$\forall s,\ \mathcal{R}_r(\hat{\gamma}_n, G_s) = \min_{\gamma \in \Omega_\gamma} \mathcal{R}_r(\gamma, G_s).$$

Then as $(\hat{\gamma}_n)_n$ is a sequence in the compact set Ω_γ there is an function $\phi : \mathbb{N} \to \mathbb{N}$ and $\tilde{\gamma}$ such that $\hat{\gamma}_{\phi(n)} \xrightarrow{d} \tilde{\gamma}$. But as Ω_γ is compact by an easy consequence of the covering lemma we get that:

$$\sup_{\gamma \in \Omega_\gamma} \left| \mathcal{R}_r(\gamma, G_s) - f(\gamma) \right| \xrightarrow{P} 0,$$

where $f : \gamma \to \lim_n \mathbb{E}(\mathcal{R}_r(\gamma, G_s))$. Therefore we have that

$$|\mathcal{R}_r(\hat{\gamma}_{\phi(n)}, G_{\phi(n)}) - f(\hat{\gamma}_{\phi(n)})| \xrightarrow{P} 0.$$

But using the expressions derived in the proof of respectively Lemma D.8 and Theorem C.1 and the hypothesis on L we have that f is continuous and is strictly convex, hence has a unique minimizer.

Therefore $\tilde{\gamma}$ must be determistic equal to $\gamma^* \triangleq \mathrm{argmin}_\gamma f(\gamma)$. Indeed suppose by contradiction that it is not the case then there is $\eta > 0$ s.t

$$P(\mathcal{R}_r(\hat{\gamma}_{\phi(s)}, G_{\phi(s)}) - \mathcal{R}_r(\gamma^*, G_{\phi(s)}) > \eta) > \eta,$$

which is a contradiction of the definition of $(\hat{\gamma}_n)_n$. Therefore we have successfully proven that $\tilde{\gamma} = \gamma^*$.

And we have proved that $\hat{\gamma}_n \xrightarrow{P} \gamma^*$.

□

References

[1] C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. *Sparse exchangeable graphs and their limits via graphon processes*. Jan. 2016. arXiv: 1601.07134.

[2] C. Borgs, J. T. Chayes, H. Cohn, and V. Veitch. *Sampling perspectives on sparse exchangeable graphs*. 2017. arXiv: 1708.03237.

[3] F. Caron and E. B. Fox. "Sparse graphs using exchangeable random measures". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.5 (2017), pp. 1295–1366. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12233.

[4] S. Chatterjee. “Concentration inequalities with exchangeable pairs (Ph.D. thesis)”. In: *ArXiv Mathematics e-prints* (July 2005). eprint: math/0507526.

[5] V. Veitch and D. M. Roy. *The Class of Random Graphs Arising from Exchangeable Random Measures*. Dec. 2015. arXiv: 1512.03099.