

DeepTrax: Embedding Graphs of Financial Transactions

C. Bayan Bruss, Anish Khazane, Jonathan Rider, Richard Serpe, Antonia Gogoglou, Keegan E. Hines
Capital One
McLean, VA 22102, USA

Abstract—Financial transactions can be considered edges in a heterogeneous graph between entities sending money and entities receiving money. For financial institutions, such a graph is likely large (with millions or billions of edges) while also sparsely connected. It becomes challenging to apply machine learning to such large and sparse graphs. Graph representation learning seeks to embed the nodes of a graph into a euclidean vector space such that graph topological properties are preserved after the transformation. In this paper, we present a novel application of representation learning to bipartite graphs of credit card transactions in order to learn embeddings of account and merchant entities. Our framework is inspired by popular approaches in graph embeddings and is trained on two internal transaction datasets. This approach yields highly effective embeddings, as quantified by link prediction AUC and F1 score. Further, the resulting entity vectors retain intuitive semantic similarity that is explored through visualizations and other qualitative analyses. Finally, we show how these embeddings can be used as features in downstream machine learning business applications such as fraud detection.

I. INTRODUCTION

Financial transactions between merchants, customers, lenders, and banks present a rich view of the economic activity within a market. It can be useful to consider this type of data as a heterogeneous graph of market participants which are connected by edges (transactions). This is a particularly useful formulation for tackling critical business problems like credit risk modeling, fraud detection, and money laundering detection. However, such a graph will be very high dimensional (with tens or hundreds of millions of vertices) and very sparse (with each vertex interacting with a fraction of the other vertices), thus limiting the graph’s utility for common machine learning tasks.

In recent years, graph embeddings techniques have grown in popularity as a means for learning latent representations of vertices on large-scale networks. Certain techniques like Graph Convolutional Networks (GCNs), DeepWalk, and node2vec attempt to encode topological structure from graphs into dense representations such that nodes with high levels of neighborhood overlap are co-located in the embedding space. This is commonly referred to as geometric similarity, which captures both graphical substructure as well as similarity among any ancillary features - e.g merchant type in the context of financial transactions - that belong to any particular vertex. Embeddings produced by the aforementioned techniques can also serve as useful features for downstream supervised tasks.

In this work, we present a novel application of graph embedding techniques to problems in financial services. In particular,

we focus on large-scale datasets of credit card transactions which define an implicit bipartite graph between account-holders and merchants.¹ We demonstrate that embedding these transactions can lead to representations which encode economic properties such as spending patterns, geography, and merchant industry/category. In Section 2, we briefly explore current literature on representation learning for large-scale graph networks. We then formally present our own method in Section 3, and quantitatively and qualitatively evaluate our results on multiple financial transaction datasets in Section 4. We conclude by demonstrating how these embeddings can benefit downstream tasks such as fraud detection.

II. RELATED WORK

A. Types of Graph Embedding Techniques

One way to group graph embedding techniques is based on the type of input they can incorporate. Inputs can be *homogeneous* where all nodes are of the same type, *heterogeneous* with multiple types of nodes and *auxiliary information* graphs that contain node, edge or neighborhood features. In homogeneous graphs, the challenge is to encode the neighborhood topology of the nodes in a computationally feasible manner [1], [2]. The latent vectors are expected to preserve different orders of node proximity (e.g. [3]) and different ranges of structural identity (e.g. [4], [5]). Therefore, the rich contextual information they carry makes node embeddings useful for multiple unsupervised learning tasks such as predicting missing links [6] as well as recommendation and ranking of the most relevant nodes [7], [8]. Furthermore, modifying the properties of random walks can assist the learned embeddings in encapsulating both local and global graph properties [9], [10], [7]. The problem of *heterogeneous* graph embedding was addressed with metapath2vec [11], where metapaths among specific entities types are defined and then random walks are generated only in accordance to those metapath schemes. This approach was extended in [12] to include node attributes and multiplex edges. Further advancements have allowed the incorporation of node and edge feature vectors (*auxiliary information*) to facilitate inductive learning of representations [13]. In these works, the estimation of node embeddings proceeds through typical sampling-based approaches [14], [15].

¹All trademarks referenced herein for illustrative and education purposes only and are owned by their respective owners.

B. Large Scale Applications In Industry

Many internet-scale recommendation systems use graph embedding techniques to supply millions of customers with potentially useful or interesting content related to their past interests [16], [17], [18], [19]. These systems typically model vertices as users, content, or products on very large graphs, and several instances of graph embeddings techniques [20] have been applied to networks with millions of unique entities, with even a few applications in the financial services space [21] using autoencoders to create embeddings on account transaction data. Embeddings from these approaches are typically used in downstream applications like product recommendation [18], [22] and maximizing proper ad placement [23]. This transfer learning approach is very similar to the impact that word embeddings have had for a variety of NLP tasks [24], [25], [26]. To our knowledge, the method proposed in this paper is the first application of graph-embeddings to financial transactions.

III. METHODOLOGY

A. Projections of Heterogeneous Graphs

The credit card transaction graph is a bipartite graph between accounts and merchants, with a transaction forming an edge. While much previous work has studied embeddings of homogeneous graphs, [11] consider heterogeneous graphs in their *metapath2vec* framework. Here, a metapath scheme is chosen to determine which sequence of node types are considered in the walks. Then, only random walks that are consistent with this scheme are generated for training embeddings. Concretely, in our bipartite graph we might consider a metapath such as $\{Account, Merchant, Account\}$ or $\{Merchant, Account, Merchant\}$. Given these schemes, we would only consider graph walks that adhere to such triplets: identifying accounts as similar because they shop at the same merchant (and vice versa). Alternatively, we can reach similar training sets by instead inducing two *homogeneous* projection graphs derived from the original bipartite graph: an accounts graph and a merchants graph. Short random walks on these homogeneous graphs will induce the same training pairs as those that would have been generated according to the short metapath schemes. As described below, this approach brings enhanced computational gains and flexibility. The results and analyses reported here will focus exclusively on the merchant embeddings.

B. Pre-processing Stage: Modeling a Graph as Pairs of Transactions

In the projection graph(s), an edge between two merchant vertices represents the presence of at least one account who made transactions at both merchants within a specified time window. The more often an account shops at the same two merchants within a fixed time window, the greater the weight on that edge. We read all transactions into a table that fits in memory, as shown in part (a) of Figure 1. We then transform this table into pairs of transactions by only keeping merchants that processed a transaction from the same account-holder within a specified time window. We set the time window

	Brand-Level Graph	Raw Merchant Graph
# nodes	$\sim 10^4$	$\sim 10^6$ raw merchants
# edges	$\sim 10^7$	$\sim 10^9$

TABLE I

DATASET DESCRIPTIONS. FOR THE BRAND-LEVEL AND RAW MERCHANT GRAPHS. NOTE THAT THE LARGER RAW MERCHANT DATASET MORE CLOSELY REFLECTS THE SIZE OF A TYPICAL TRANSACTION GRAPH.

depending on the node type that we are looking at as well as the density of the graph. For instance we would want a smaller window for grouping similar accounts on a merchant than we would for grouping merchants on an account. In general, we find that increasing the time window too large can lead to significantly more connections between unrelated merchants, which decreases the quality of the embeddings.

Storing this information within a table allows the user to distribute the aforementioned time windowing strategy over all rows of merchants, instead of individually running random walk operations from every merchant vertex². Part (b) in Figure 1 shows examples of transaction pairs after time-windowing. Multiple transaction pairs with the same two merchants indirectly represent a weighted edge between those merchants on a graph. In effect, this formulation allows us to model a weighted graph without explicitly creating one.

1) *Brand Level Merchant Names vs Raw Merchant Names:* When creating training pairs, we consider two different approaches. The first is using the raw merchant name and appending it with the zip code. The raw merchant name differentiates between franchise locations, but some unrelated merchants have the same raw merchant name. For this reason, the zip code is appended to the name to properly differentiate. This gives us many merchants to work with, as well as a less dense graph as accounts are less likely to be paired together. Table I shows that the raw merchant graph contains approximately 10^6 merchants and 10^9 edges.

The other approach is to use the brand-level merchant name which rolls up all franchises to the same name. This creates a highly interconnected graph, as everyone nationwide who shops at a particular brand is likely to be paired with another similar shopper. This causes the number of pairs formed with the brand-level graph to be far larger than those of the raw merchant. Additionally, for rarely-occurring merchants, it can be challenging to accurately identify the correct brand entity. Due to these factors, we drop any merchants with fewer than 50 transactions per day. Table I shows that the brand-level graph contains approximately 10^4 merchants and 10^7 edges.

2) *Separating Online from Offline Transaction Pairs:* Prior to training, we place online merchants and physical merchants into two distinct tables. When looking at raw merchant names we find that online retailers - e.g *Amazon.com*, *Newegg.com* - often precede or follow several other unrelated merchants in transaction sequences. Even though there are physical merchants that also frequently appear in many transaction pairs, these vendors are typically separated into several stores with distinguishable identifiers - e.g *McDonalds 94123*. Con-

²While algorithms for distributed random walks on graphs do exist, these techniques are non-trivial to implement and excessive for creating pairs of transactions (or equivalently, short walks of length 2).

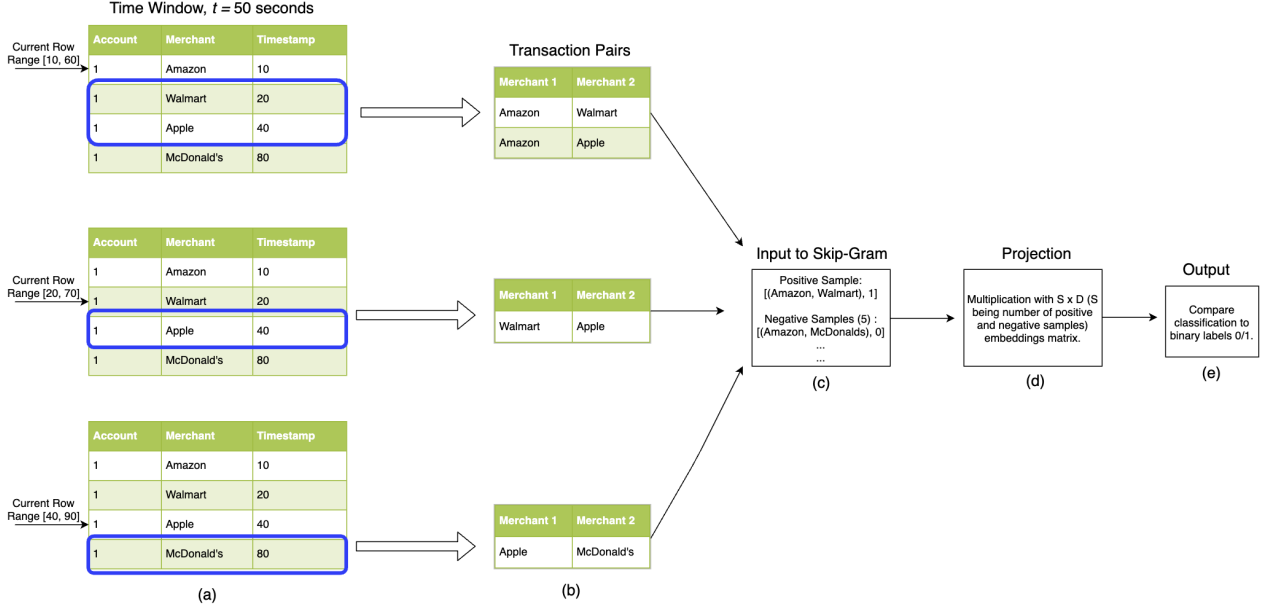


Fig. 1. Model pipeline, from data pre-processing to training with Skip-gram. Using stringent time windows and pairs of transaction pairs (example time window, $t = 50$ seconds) allows for creating meaningful embeddings on graphs with millions of unique entities.

sequently, each store is likely to be close to other merchants in the same geographic region. Online merchants, however, are location-agnostic and similar to supernodes on a graph that have a disproportionately high number of edges. Training on each set separately allows us to create embedding representations for merchants that are not biased by artificial relationships in the raw transactional data.

C. Approximating DeepWalk with Transaction Pairs

We posit that for this financial graph, short-range interactions (one-hop and two-hop neighborhoods) will be sufficient to yield effective embeddings. When viewed in this way, we can consider random walks on a graph in the limit of either (i) very short walk lengths or (ii) very short context windows within walks. At this short-range limit, we consider only two-hop walks on the bipartite graph, or equivalently, one-hop walks on the homogeneous projection. Our embeddings are trained using negative sampling, with node-pairs generated as just described used as positive samples. Negative samples are generated by sampling nodes at random (assuring no actual edge exists), using the negative sampling strategy described in [27]. This leads to optimization of the following loss:

$$\begin{aligned} \underset{\phi}{\text{minimize}} \mathcal{L}(\phi) = & - \sum_{m \in V} [\log P(y = 1 | \phi(m), \phi(m_{pos})) \\ & + kE[\log P(y = 0 | \phi(m), \phi(m_{neg}))]], \quad (1) \end{aligned}$$

where $m \in V$ denotes a merchant selected from a set of unique merchant entities, V , the mapping function $\phi : m \in V \mapsto \mathbb{R}^d$ retrieves the embedding representation for any given merchant, m_{pos} denotes a positive sample merchant for a given

merchant, and m_{neg} denotes a negative sample chosen from V and k is the number of negative samples.

If we interpret equation (1) from a graph-based perspective, minimizing this objective function amounts to creating embedding representations that capture first-order proximity relationships between different merchants on a graph. Representing the transactional graph as pairs forces the model to capture these first-order relationships. Part of the motivation for not exploring higher-order relationships is to guarantee true noise samples during training. As transactional data is often noisy, maintaining a small context window makes it easier to guarantee that a negatively sampled merchant does not appear in the immediate vicinity of a merchant of interest. A larger context window also increases the probability of interrelating merchants that are not actually meaningfully similar. However, the use of the time-window strategy does allow for a tuneable parameter that can act similarly to capturing higher order relationships. Intuitively, if positive pairs are generated from all merchants that a single account has shopped at in an entire month, then this is a higher order proximity than if only a one hour time window had been considered.

Furthermore, [28] demonstrates that if the number of random walks per vertex is large enough, the expected average walk length for each one of them will converge to the shortest path between source and destination vertex. In other words, by significantly increasing the number of walks, the expected walk length reduces to that of the shortest path. Thus, using transaction pairs - or truncated random walks of length 2 - serves as an approximation to the shortest path between a merchant and every semantically similar merchant. Our qualitative and quantitative analysis in Section IV demonstrates that this approximation is effective.

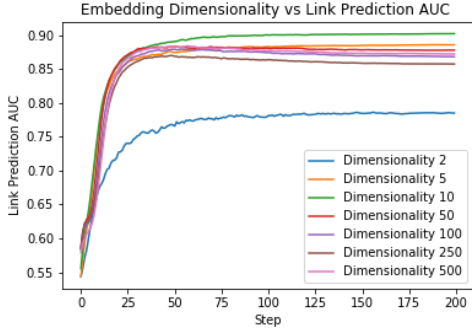


Fig. 2. Impact of embedding dimensionality. While increasing embedding dimensionality yields very large quantitative improvements on LPA at low dimensionality, results quickly stagnate after 10 or more dimensions and overfitting is observed.

	Link Prediction AUC (LPA)	F1 Score
Brand-Level Merchant Graph	0.90	0.67
Raw Merchant Graph	0.68	0.67

TABLE II

MODEL PERFORMANCE AFTER 2 EPOCHS OF TRAINING ON ALL DATASETS. THE MODEL LIKELY PERFORMS VERY WELL ON THE SMALLER BRANDED DATASET DUE TO FEWER INSTANCES OF OVERREPRESENTED / UNDERREPRESENTED MERCHANTS IN THIS GRAPH.

IV. RESULTS

We train our model on both merchant datasets for roughly 2 epochs on 48 vCPUs. For quantitative analysis, we present F1, link prediction AUC and area under the precision-recall curve for evaluating our embeddings with an internal fraud detection tool. We report F1 and LPA scores from running experiments on both the raw and brand-level merchant datasets, but use the brand-level dataset for presenting embeddings visualizations due to its smaller size.

A. Performance on Brand-Level Network

1) *Quantitative Analysis*: We start by analyzing our model’s effectiveness on the brand-level graph. In Table II, we see that the model achieves a link prediction AUC of roughly 0.91 and a F1 score of 0.67. The model’s high performance on both metrics demonstrate its capability to learn meaningful relationships even within smaller-sized graphs of roughly 10,000 or fewer elements. This performance also demonstrates that training on pairs of transactions (as opposed to long sequences) is not detrimental to creating high-quality embeddings. This is to be expected, as training Skip-gram with a sequence length of 2 simplifies its objective function to optimize at every iteration. This in turn forces the system to create node embeddings that encode meaningful information about the edges in their surrounding subgraph.

In Figure 2, we present an analysis on how embedding dimensionality affects the model’s performance. We see that increasing dimensionality from 2 to 5 yields a roughly 13% improvement in link prediction AUC, but further increasing this hyperparameter does not yield significant benefits on this particular metric. This is likely due to the brand-level dataset’s small vocabulary size (roughly 10^4 unique merchants). This diagram suggests that only a small embedding dimensionality

Adidas	Delta Air Lines	West Elm
Nike	United Airlines	Anthropologie
Reebok	American Airlines	Crate&Barrel
Timberland	Frontier Airlines	Pottery Barn
KFC	Starbucks	Gap
Taco Bell	Panera Bread	Banana Republic
Little Caesars	Dunkin Donuts	Ann Taylor Loft
Burger King	Jamba Juice	J Crew

TABLE III
MERCHANT NEAREST NEIGHBORS DERIVED FROM THE BRAND-LEVEL MERCHANT GRAPH. FOR EACH QUERY MERCHANT VECTOR (IN BOLD), THE TOP THREE NEAREST NEIGHBOR MERCHANT VECTORS ARE SHOWN.

is required to adequately capture semantic similarity for the brand-level transactional dataset discussed in this paper.

2) *Qualitative Analysis*: The brand-level merchant graph embeddings provide intuitive consistency. With word embeddings, a common observation is that words which are semantically similar tend to be embedded in close proximity. Here, we redefine semantic similarity to be sets of merchants which are interchangeable for any given commerce purpose. That is, two merchants are semantically similar if they exist in the same industry, category, price point or all of the above. With this in mind, we report in Table III the nearest neighbor merchant vectors for several household brands. For example, we see that the nearest neighbors to the **KFC** vector are not only other restaurants, but other fast food competitors. As can be seen in Table III, this holds true across many industries including fashion, air travel, food, and furniture. The *West Elm* vector presents an interesting result. Two of the three nearest neighbors are obviously correct: *Pottery Barn* and *Crate&Barrel* are also furniture manufacturers. However, the closest neighbor in the entire merchant space is actually *Anthropologie*, a store which is most commonly known as a fashion brand. However, within *Anthropologie*’s offerings is an extensive home furnishings and furniture section. Overall, these findings indicate that semantically similar merchants are typically embedded close together.

This general pattern of separability-by-industry extends across the whole embedding space. Figure 3 shows a low-dimensional visualization of the embedding space for our brand-level merchant graph. As expected, merchants which serve the same industry or category tend to co-locate in similar areas of the embedding space: a visual extension of the results of Table III. In the left of Figure 3, we can note that sporting goods brands such as *Nike*, *Under Armour*, and *Columbia* are located close to each other. Similarly, note that airlines are co-located, as well as fast food (top). Finally, the bottom close-up shows a region of the embedding space with merchants such as *Jpay*, *INMATE PAYMENT*, *SECURUS INMATE CALL-V*. These companies provide a set of services for telecommunications and payments into and out-of the American prison system. That is, the customers of these companies are inmates (and family members of inmates) who must use these merchants to conduct common activities. Due to these specialized services, there should be very little overlap between these merchants and any accounts not affiliated with inmates. It is encouraging that our graph embedding system is able to accurately embed these merchants together.

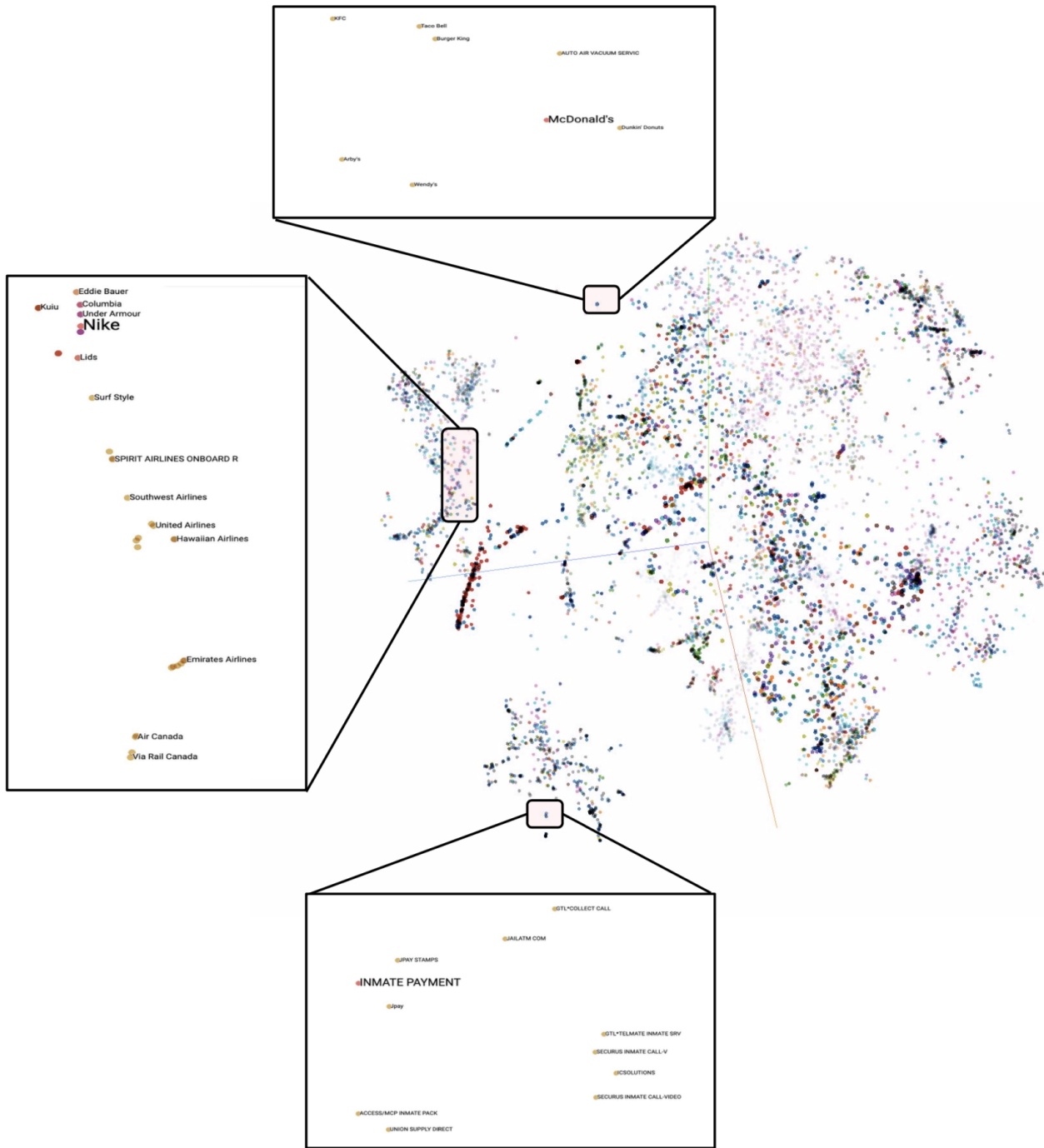


Fig. 3. Example clusters found in a two dimensional t-SNE projection of brand-level embeddings. Best viewed digitally.

A final set of questions we can ask with these brand-level merchant embeddings is whether relationships between merchants can be consistently observed within and between industries. In much the same way that analogical reasoning tasks can be accomplished with word vectors, it might be feasible to identify compelling relationships between merchant vectors.

One way to construct analogies for merchants is to devise relations between merchants within one category and determine if the same relationship holds (geometrically) across categories. As a concrete example, we can recognize that

within a particular industry, there will exist several offerings that are not direct competitors but instead operate at different price points. For example, within automobiles, the typical *Porsche* price point is well above that of *Honda*, though they offer the same basic good. If there is a direction in the embedding space that encodes price point (or *quality* or *luxury*) then this component should contribute to all merchants across all industries. In this way, the embeddings can elucidate analogies such as "Porsche is to Honda as *blank* is to Old Navy".

Uncovering such a direction can be achieved in several ways

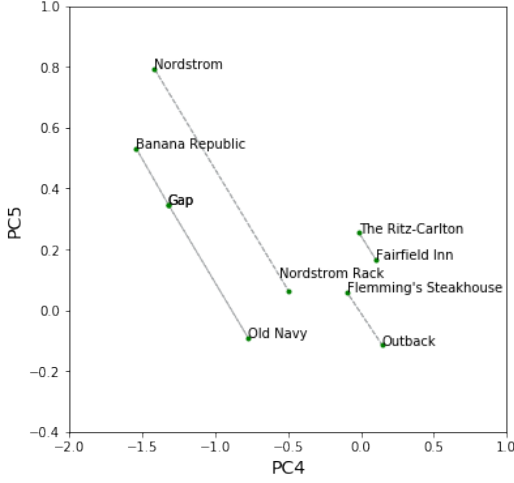


Fig. 4. Merchant embeddings tend to encode typical price point of goods. Merchants from multiple industries visualized in the subspace spanned by the fourth and fifth principal components. Pairs of merchants are joined by dotted lines, with each pair containing a high-end merchant and a more affordable counterpart. Across disparate industries, the relationship between high-end offerings and affordable offerings is embedded in a consistent direction of this space.

([29], [30]). For our purposes, we assumed that such a direction, if it existed, was likely a dominant source of variation between the embeddings and that this direction is likely captured by one or a few principal components of variation. With this in mind, we identified several pairs of merchants which exist in the same category yet span a range of price points. These included: {Gap and Old Navy}, {Nordstrom and Nordstrom Rack}, {Ritz-Carlton and Fairfield Inn}, and so on. These pairs were chosen because they represent nearly identical offerings within a category, but at a high and low price point. In Figure 4, we visualize these paired merchant vectors as projected into a convenient subspace spanned by two principal components. Interestingly, the relationship (slope) between the low-end and high-end offering is nearly parallel for all of these pairs. There indeed exists a *direction* in the embedding space which encodes price point and this direction is consistent across these disparate industries.

B. Performance on Raw Merchant Network

	% Δ in AUpr
Fraud Detection Tool + Raw Merchant Embeddings	+0.9%
Fraud Detection Tool + MLP[Raw Merchant Embeddings]	+5.2%

TABLE IV

TRAINING AN INTERNAL FRAUD DETECTION TOOL WITH ONLY MERCHANT EMBEDDINGS OR WITH PROJECTED MERCHANT EMBEDDINGS (THROUGH A MULTI-LAYER PERCEPTRON) YIELDS HIGHER CLASSIFICATION PERFORMANCE OVER AN INTERNAL BASELINE MODEL.

1) *Quantitative Analysis*: We see in Table II that the model performs similarly on the raw merchant graph with respect to F1 score. Achieving a 0.67 score on the one million plus merchant graph demonstrates the model’s ability to maintain

meaningful embedding representations that are not heavily influenced by outliers in the larger-sized graph. While the model scores lower on lower link prediction AUC (0.68) for this dataset, we expect to see a drop-off due to training on a sparser graph; there are far more instances of merchants in this network that are only connected to a few neighboring vendors. Still, even with this obstacle, the model is able to create embedding vectors that are semantically meaningful as indicated by the top-5 closest neighbors to the examples given in Table V. For example, the first two columns show fast-food merchants that not cluster together by type - e.g *Dunkin*, *Nature’s Way Cafe* - but also by geographic region.

2) *Qualitative Analysis*: Table V shows nearest neighbors for several raw merchant vectors. Since these raw merchant entities correspond to physical locations, it is not surprising that geography plays a large role in this embedding space. In Table V, each entity is shown alongside its zipcode, confirming that merchant vectors tend to be embedded near other merchants in the same geographic area. (Note that zipcode and geography are not inputs to the model). We see that the top-5 neighbors for *DUNKIN #332240 Q35* are not only other stores in the same chain - e.g *DUNKIN #341663*, *DUNKIN #341663 Q35* - but even loosely related cafes like *NATURE’S WAY CAFE BO* that lie within neighboring counties in Florida. Further, the nearest neighbors here highlight what is potentially a naming or logging error in the point-of-sale system. Note that the same physical store location *DUNKIN #341663* shows up under two entity names: *DUNKIN #341663 Q* and *DUNKIN #341663 Q35*. This demonstrates an interesting potential for this kind of analysis to be leveraged for entity resolution based not on string similarity, but based on correlated shoppers. Finally, the second and third examples illustrate some ways that the model can capture local geographic cultural nuances such as the high proportion of breweries in Portland, Oregon where Powell’s Burnside is located, when compared the number of coffee shops in Los Angeles, CA where The Last Book Store is located.

DUNKIN #332240 Q35	33442	POWELL’S BURNSIDE	97209	THE LAST BOOK STORE	90013
CHUCK E CHEESE 682	33428	TRIMET TVM	97202	PAMPAS GRILL- STYL	90036
DUNKIN #341663 Q	33442	TARGET 00027912	97205	SQ *BLUE BOTTLE COF	90013
DUNKIN #341663 Q35	33442	10 BARREL BREWING CO	97209	HIGHLAND PARK BOWL	90042
NATURE’S WAY CAFE BO	33431	DESCHUTES BREWERY	97209	VERVE COFFEE ROASTERS	90014
DD/BR #338392 Q3	33073	PORTLAND JAPANESE GARD	97205	GRILL CONCEPTS - S	90404

TABLE V

EXAMPLES OF MERCHANT SIMILARITIES WHEN TRAINED USING THE RAW MERCHANT NAME. EACH MERCHANT IS SHOWN ALONGSIDE ITS ZIP CODE. GEOGRAPHY IS A STRONG SIGNAL IN THESE EMBEDDINGS, BUT SEMANTIC AND REGIONAL INFLUENCE CAN ALSO BE OBSERVED.

C. Application to Fraud Detection

We also assess the quality of our raw merchant embeddings by evaluating them in a transfer learning task involving transaction fraud detection. Results on these experiments are reported relative to a baseline model (the details of which we omit here) and are quantified by the area under the precision-recall curve (AUpr). In Table IV, we see that directly using the trained embeddings from the raw merchant graph yields a roughly 1.0% improvement in fraud classification AUpr when these embeddings are included as additional features to the

model. One downside of this approach is the large expansion of the feature space in order to include the embeddings. To overcome this, we additionally tested a model whereby we added a trainable MLP that took as input the embedding for a transaction's merchant and predicted a binary output for fraud. This ancillary model, once trained, can then be used to output a single fraud score per transaction, conveying only the information contained in the merchant embedding space that is useful for fraud detection. This single score is then passed into the base model as an additional feature and yields a 5.2% boost in efficacy. Merchants that engage in fraudulent transactions typically engage with similar vendors over all transaction pairs. As our model's objective is to encode transactional behavior within each merchant's embedding representation, we attribute an increase in classification accuracy to capturing semantically meaningful features that allow the classifier to better identify likely-fraudulent merchants.

V. CONCLUSION

In this paper, we propose an approach for training embeddings of entities from financial transactions. Our approach poses sequences of financial transactions as a graph, where a customer engaging in a transaction at two merchants within a specified time window constitutes an edge between those two merchant in the network. We demonstrate that this approach results in semantically meaningful embedding vectors for up to millions of unique merchant entities. We quantitatively show in Section IV that our model scores strongly with respect to link prediction AUC and F1 evaluation scores, and also provides lift in classification accuracy for an internal fraud detection tool.

The results presented here were primarily based on capturing network topology only, while omitting ancillary attributes about accounts, merchants, and transactions. Future work remains to be done to incorporate techniques such as those described in ([6], [12]). Lastly, we hope to analyze how embeddings impact other downstream applications in the financial services such as marketing and credit charge-off prediction.

REFERENCES

- [1] Y. Zhang, T. Lyu, and Y. Zhang, "COSINE: community-preserving social network embedding from information diffusion cascades," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 2620–2627.
- [2] C. Zhou, Y. Liu, X. Liu, Z. Liu, and J. Gao, "Scalable graph embedding for asymmetric proximity," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017, pp. 2942–2948.
- [3] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," *CoRR*, vol. abs/1403.6652, 2014.
- [4] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 385–394. [Online]. Available: <https://doi.org/10.1145/3097983.3098061>

- [5] S. Abu-El-Hajja, B. Perozzi, R. Al-Rfou, and A. A. Alemi, "Watch your step: Learning node embeddings via graph attention," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montr al, Canada.*, 2018, pp. 9198–9208.
- [6] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," *CoRR*, vol. abs/1607.00653, 2016. [Online]. Available: <http://arxiv.org/abs/1607.00653>
- [7] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1225–1234.
- [8] J. Weston, S. Chopra, and K. Adams, "#tagspace: Semantic embeddings from hashtags," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1822–1827.
- [9] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, 2015, pp. 891–900.
- [10] B. Perozzi, V. Kulkarni, H. Chen, and S. Skiena, "Don't walk, skip!: Online learning of multi-scale network embeddings," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017*, 2017, pp. 258–265.
- [11] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: ACM, 2017, pp. 135–144. [Online]. Available: <http://doi.acm.org/10.1145/3097983.3098036>
- [12] Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou, and J. Tang, "Representation learning for attributed multiplex heterogeneous network," 2019.
- [13] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *CoRR*, vol. abs/1706.02216, 2017. [Online]. Available: <http://arxiv.org/abs/1706.02216>
- [14] M. Gutmann and A. Hyv rinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," *Proceedings of Machine Learning Research*, vol. 9, pp. 297–304, 13–15 May 2010. [Online]. Available: <http://proceedings.mlr.press/v9/gutmann10a.html>
- [15] M. U. Gutmann and A. Hyv rinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 307–361, Feb. 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2503308.2188396>
- [16] M. Grbovic and H. Cheng, "Real-time personalization using embeddings for search ranking at airbnb," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 311–320.
- [17] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich, "Pytorch-biggraph: A large-scale graph embedding system," in *Proceedings of the 2nd SysML Conference*, 2019.
- [18] J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee, "Billion-scale commodity embedding for e-commerce recommendation in alibaba," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 839–848.
- [19] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 974–983.
- [20] H. Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1416–1424.
- [21] L. Baldassini and J. A. R. Serrano, "client2vec: Towards systematic baselines for banking applications," *CoRR*, vol. abs/1802.04198, 2018. [Online]. Available: <http://arxiv.org/abs/1802.04198>
- [22] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 974–983.
- [23] E. Ordentlich, L. Yang, A. Feng, P. Cnudde, M. Grbovic, N. Djuric, V. Radosavljevic, and G. Owens, "Network-efficient distributed

- word2vec training system for large vocabularies,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 1139–1148.
- [24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: <http://arxiv.org/abs/1310.4546>
 - [25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
 - [26] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *CoRR*, vol. abs/1906.08237, 2019. [Online]. Available: <http://arxiv.org/abs/1906.08237>
 - [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
 - [28] Z. Zheng, H. Wang, S. Gao, and G. Wang, “Comparison of multiple random walks strategies for searching networks,” *Mathematical Problems in Engineering*, vol. 2013, 2013.
 - [29] O. Levy and Y. Goldberg, “Linguistic regularities in sparse and explicit word representations,” in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, 2014, pp. 171–180.
 - [30] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, 2013*, pp. 746–751.