COS424 Assignment 3: Bitcoin

Sagar Setru
Quantitative and Computational Biology
sagar.setru@princeton.edu

Hugh Wilson
Quantitative and Computational Biology
hswilson@princeton.edu

Abstract

We predicted transactions between Bitcoin users using low dimensional representations of previous transactions, generated from two approaches: one, graphical methods and two, matrix factorization followed by fitting of a Gaussian Mixture Model. We found that the undirected graphical features produced higher AUROC scores (0.662 - 0.764) than directed graphical features in general, but that the directed feature sender pagerank (with an AUC score of 0.666) maintained a higher precision (0.6 compared with < 0.4 for the undirected features) for recall values up to 0.4. The AUC score for the matrix factorization method with clustering across the senders (0.705) was comparable to using the undirected graphical features.

1 Introduction and Related Work

Bitcoin is a virtual currency that guarantees user anonymity. However, the Bitcoin blockchain is a publically available ledger of the transaction history between all Bitcoin users, who are given anonymized addresses. A blockchain can be represented as a graphical model, with nodes for user addresses and edges for transactions between users; or, as a sparse interaction matrix between addresses. Despite the anonymity inherent to the blockchain, the data it contains may be mined to predict future transactions [Mei+13].

Several methods exist for characterizing nodes in directed graphs and pairs of nodes in undirected graphs, and these methods have been shown to be useful for predicting future interactions between nodes [Mur12; Pag+99; Kle99; LM05; LK07; ZLZ09]. For directed graphs, these methods include the PageRank algorithm, which has at its heart a model of a directed graph as the stationary distribution of a Markov chain and is most notably used as a part of the Google search algorithm to rank the webpages that match a given query [Pag+99; LM05; Mur12]. Another graphical feature extraction method includes the Hypertext Induced Topic Search (HITS) algorithm, an iterative algorithm that characterizes each node with a hub value estimated based on outgoing links and an authority value estimated based on incoming links [Kle99; LM05]. For undirected graphs, there exist several methods for generating features to predict links between pairs of nodes u, v based on the set of neighbors for each node $\Gamma(u), \Gamma(v)$. These include the resource allocation index [ZLZ09], preferential attachment score [LK07], Adamic-Adar index [AA03; LK07], and Jaccard coefficient [LK07].

Besides the graphical representation of the blockchain, one can also consider it as a sparse interaction matrix that contains latent, low dimensional structure that can be extracted using matrix factorization (MF) methods. MF methods produce a low dimensional representation of data $z_i \in \mathbb{R}^L$, which can then be clustered, for example using a mixture model. Several MF methods exist for producing low dimensional representations of high dimensional data, including principal components analysis, probabilistic MF, nonnegative MF, Poisson factorization, and singular value decomposition.

Motivated by the various MF methods as well as those methods for characterizing nodes in directed and undirected graphs, we sought to evaluate how well these two broad classes of methods predict future transactions between Bitcoin users from previous transactions.

2 Methods

Two datasets were used, one for training and the other for testing. The training dataset is the number of transactions between Bitcoin users (sender, receiver pairs) during one year. We assumed that a zero value indicated there were no transactions between that sender and receiver. (It should be noted that the absence of a Bitcoin transaction does not mean the absence of any type of financial transaction - bank transfer, cash exchange...etc. However, this other transaction data is missing based on a separate observed variable - the transaction type - and so is missing at random.) The testing dataset lists a number of sender, receiver pairs and whether a transaction occurred in the year following the training year.

Graphical features were generated by treating the training dataset as an interaction matrix between nodes on a graph. Edges between nodes were labeled with the number of transactions between the nodes, normalized by the total number of outgoing transactions, such that $\Sigma_{j=1}^{N} A_{ij} = 1$, where A is the normalized interaction matrix and A_{ij} is the number of transactions between sender i and receiver j. The graphical features we generated can be broken up into two sets: directed and undirected. The directed features were generated from a transaction weighted graph in which the edge direction is from sender to receiver. They include the hub (H), authority (A) [Kle99; LM05], and PageRank (PR) [Pag+99; LM05; Mur12] values for each node, calculated using the HITS and PageRank algorithms, respectively. The undirected features were calculated by ignoring the edge directionality.

For pairs of nodes u, v, define the sets of neighbors for each node as $\Gamma(u), \Gamma(v)$ and the number of elements in the set $\Gamma(u)$ as $|\Gamma(u)|$. In these terms, undirected features include the resource allocation index (RAI) $\Sigma_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|}$ [ZLZ09], preferential attachment score (PAS) $|\Gamma(u)| \cdot |\Gamma(v)|$ [LK07], Adamic-Adar index (AAI) $\Sigma_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$ [AA03; LK07], and Jaccard coefficient (JC) $\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$ [LK07]. The undirected and directed networks were instantiated and their respective features calculated using the NetworkX Python libraries, with the default parameters for all methods [HSS08]. The directed features were calculated for each node, giving two values per sender-receiver pair, whereas the undirected features were calculated for each pair of nodes, giving one value per sender-receiver pair.

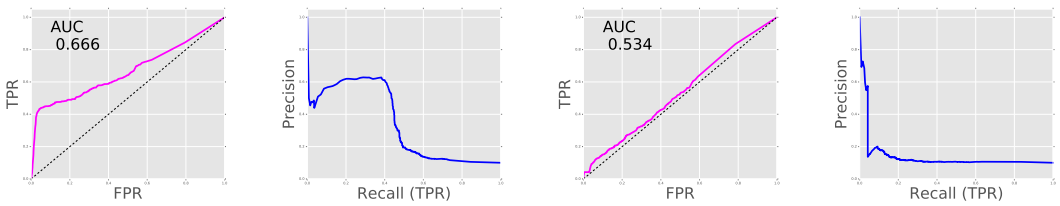
Matrix factorization was used to produce a low-dimensional representations of the data. We considered probabilistic matrix factorization (PMF)[SM], nonnegative matrix factorization (NMF)[LS99], Poisson factorization (PF)[Gop+], hierarchical PF (HPF)[GHB], principal components analysis (PCA), Kernel PCA (KPCA), singular value decomposition (SVD), and truncated SVD (TSVD). The low-dimensional features generated by the MF methods were then clustered using a Gaussian Mixture Model (GMM) with two components. Due to computational constraints, we were only able to implement the TSVD method, with which we projected data into low-dimensional spaces of dimensionality $n = 1, 2, 3, 4, 5, 10, 20$. MF and GMM clustering were implemented using the SciKitLearn Python libraries [Ped+11].

The performance of all methods was determined using receiver operating characteristic curves (ROCC); precision-recall curves (PRC); and the area under the ROC curve (AUC). These metrics were produced using the SciKitLearn Python libraries [Ped+11].

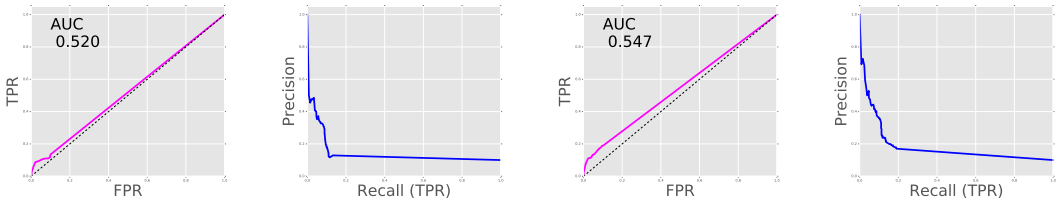
3 Results

ROC and precision-recall curves were plotted for each of the directed (Figure 1) and undirected (Figure 2 graphical features. The ROC and precision-recall curves provide complementary information. The ROC curve shows the fraction of ground truths identified as true against the fraction of ground falsehoods identified as true. The precision-recall curve shows the fraction of predictions which are correct against the fraction of ground truths identified as true. The precision-recall curve is useful to consider when the probability of a ground truth is much lower than that of a ground falsehood, as here.

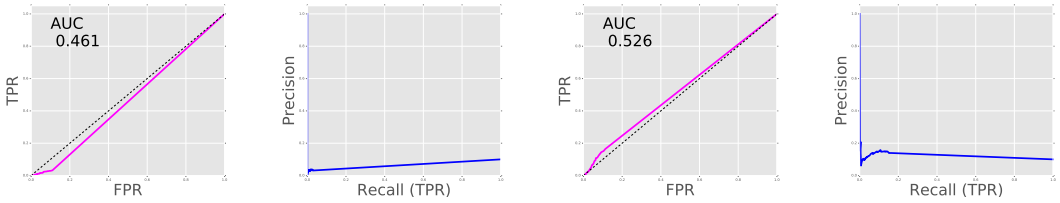
The AUC scores for the undirected features (Figure 2 were within the range: 0.662 - 0.764, with the highest score corresponding to preferential attachment (PAS). The PR curves show low values of precision (0 - 0.4) with a tendency to decrease as a function of recall. The AUC scores for the directed features (Figures 1 were within 0.05 of 0.5 for all features except the sender page rank, which had a score of 0.666 - comparable to the lower end of the undirected feature range AUC scores. The PR curves show low precision values (0 - 0.2), except for sender pagerank which maintains a high precision, 0.6, up to a recall value of 0.4. Predictions were also made using the mean of all of the graphical features. The curves produced were similar to those for PAS (see Appendix).



(a) ROCC, Sender PR (b) PRC, Sender PR (c) ROCC, Receiver PR (d) PRC, Receiver PR

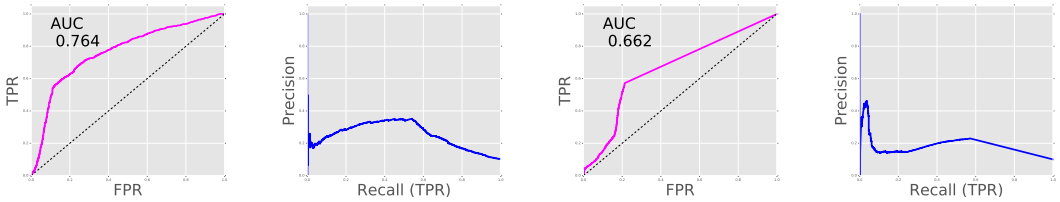


(e) ROCC, Sender A (f) PRC, Sender A (g) ROCC, Receiver A (h) PRC, Receiver A

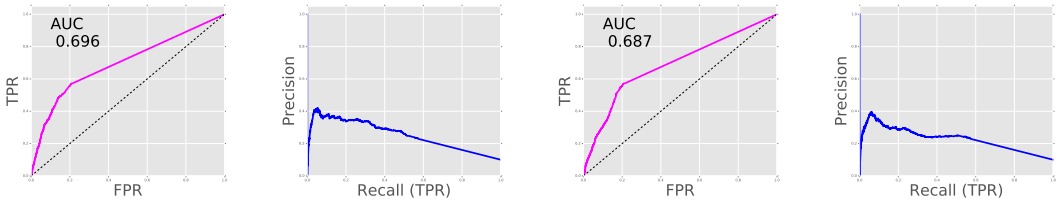


(i) ROCC, Sender H (j) PRC, Sender H (k) ROCC, Receiver H (l) PRC, Receiver H

Figure 1: ROCC and PRC when predicting future Bitcoin transactions using the directed features PR (a-d), A (e-h), and H (i-l), where each feature may correspond to the senders or receivers in the pairs of users from the test set.



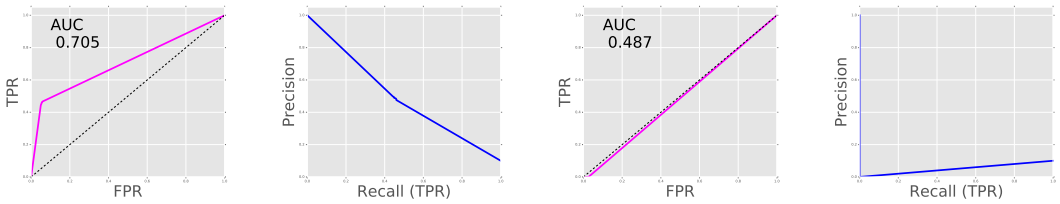
(a) ROCC, PAS (b) PRC, PAS (c) ROCC, JC (d) PRC, JC



(e) ROCC, AAI (f) PRC, AAI (g) ROCC, RAI (h) PRC, RAI

Figure 2: ROCC and PRC when predicting future Bitcoin transactions using the undirected features PAS (a-b), JC (c-d), AAI (e-f), and RAI (g-h), where each feature corresponds to the pairs of senders and receivers in the test set.

ROC and precision-recall curves were also plotted for the matrix factorization method using a 5 dimensional low dimensional space - which is representative of the results using low dimensional representations with one to ten dimensions(Figure 3). For clustering across the senders the AUC is comparable to the undirected graphical methods



(a) ROCC Sender (b) PRC Sender (c) ROCC Receiver (d) PRC Receiver

Figure 3: ROCC and PRC when predicting future Bitcoin transactions using a two component Gaussian Mixture Model fitted to low dimensional representation of the sender (a-b) and receiver (c-d) data. In each case the low dimensional space had dimension five and was produced using TrucatedSVD.

(0.705) and the PR curve shows high precision at low recall. For clustering across the receivers, the AUC is close to 0.5 (0.487) and the precision is low for all recall values.

4 Discussion and Conclusion

The AUC scores of the undirected features were larger than 0.662. This suggests the undirected features give low dimensional representations of the data that have power for predicting transactions. Most directed edges have AUC scores 0.5. This suggests that they lack predictive power. The exception is the sender PR, with AUC = 0.666. Although this is at the lower end of the AUC score range for undirected features, the PRC for sender PR maintains a higher precision than the undirected features (0.6 compared to < 0.4) for recall values < 0.4. Therefore, the sender PR might be a better method to use in cases where it is important that a high fraction of the predictions are correct and it is not important that many ground true transactions are missed. For instance, this would be helpful if someone were making bets that pay out when a predicted transaction occurs.

An alternative for situations which require high precision and accept low recall is the matrix factorization and clustering method across senders. This has an AUC comparable to the undirected graph methods (0.705) and a PRC with high precision at low recall. The success of this method suggests that is it not necessary for low dimensional representations to be intuitively interpretable (as in the graphical case, for example) in order to have predictive power. The poorer predictions from clustering across receivers are surprising, given that the identity of the receiver is likely to be important in whether a transaction will occur.

To conclude, we used two methods to extract a low dimensional representation from one year of Bitcoin transaction data in order to predict transactions between pairs of users in the following year. In the first method, we produced a graph with edges as transactions and calculated 10 features as candidate low dimensional representations of the data. In the second method, MF was used to give low dimensional representations of the transactions between senders and receivers and then a GMM was fit (separately) to the senders and receivers of the test set to make predictions.

We analyzed the capacity of each method to correctly predict transactions in the following year using ROCC and PRC. We found that the undirected features produced higher AUC (0.662 - 0.764) than directed features in general, but that the directed feature sender PR, AUC = 0.666, maintained a higher precision score (0.6 compared with < 0.4 for the undirected features) for recall values < 0.4. The AUC for MF with GMM clustering across the senders (AUC = 0.705) was comparable to using the undirected graphical features.

There several possible extensions to this work. Additional methods for characterizing pairs of nodes in undirected graphs incorporate so-called "community information" alongside local information. These methods, such as the community-based common neighbors and resource allocation algorithms developed by Soundarajan and Hopcroft [SH12], have been shown to be more strongly predictive of links in some networks. A future investigation may want to examine how to properly define a Bitcoin user's community and then use community-based methods to make predictions of Bitcoin transactions. It would also be useful to have the computational resources to work more freely with the full interaction matrix. This would allow use of more MF algorithms and clustering methods which work directly with the interaction matrix.

References

[AA03] Lada A Adamic and Eytan Adar. "Friends and neighbors on the web". In: *Social networks* 25.3 (2003), pp. 211–230 (cit. on pp. 1, 2).

[GHB] Prem Gopalan, Jake M. Hofman, and David M. Blei. "Scalable Recommendation with Hierarchical Poisson Factorization". In: UAI2015. URL: http://auai.org/uai2015/proceedings/papers/208.pdf (cit. on p. 2).

[Gop+] Prem Gopalan et al. "Bayesian Nonparametric Poisson Factorization for Recommendation Systems". In: 17th AISTATS. URL: https://www.cs.princeton.edu/~rajeshr/papers/GopalanRuizRanganathBlei2014.pdf (cit. on p. 2).

[HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring network structure, dynamics, and function using NetworkX". In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, Aug. 2008, pp. 11–15 (cit. on p. 2).

[Kle99] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632 (cit. on pp. 1, 2).

[LK07] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American society for information science and technology* 58.7 (2007), pp. 1019–1031 (cit. on pp. 1, 2).

[LM05] Amy N Langville and Carl D Meyer. "A survey of eigenvector methods for web information retrieval". In: *SIAM review* 47.1 (2005), pp. 135–161 (cit. on pp. 1, 2).

[LS99] Daniel D. Lee and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (Oct. 21, 1999), pp. 788–791. URL: http://dx.doi.org/10.1038/44565 (cit. on p. 2).

[Mei+13] Sarah Meiklejohn et al. "A fistful of bitcoins: characterizing payments among men with no names". In: *Proceedings of the 2013 conference on Internet measurement conference*. ACM. 2013, pp. 127–140 (cit. on p. 1).

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029 (cit. on pp. 1, 2).

[Pag+99] Lawrence Page et al. "The PageRank citation ranking: bringing order to the web." In: (1999) (cit. on pp. 1, 2).

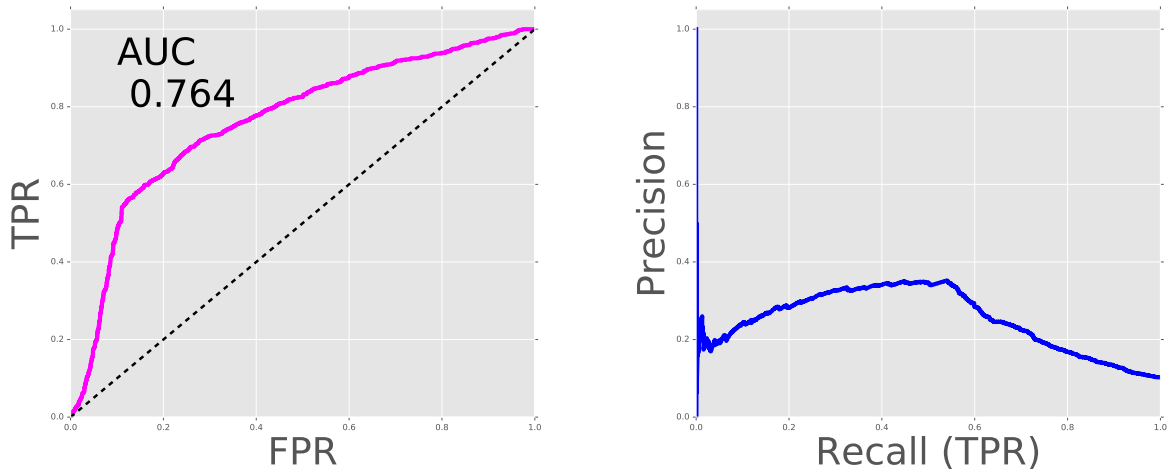
[Ped+11] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 2).

[SH12] Sucheta Soundarajan and John Hopcroft. "Using community information to improve the precision of link prediction methods". In: *Proceedings of the 21st international conference companion on World Wide Web*. ACM. 2012, pp. 607–608 (cit. on p. 4).

[SM] Ruslan Salakhutdinov and Andriy Mnih. "Probabilistic Matrix Factorization". In: NIPS. URL: https://papers.nips.cc/paper/3208-probabilistic-matrix-factorization.pdf (cit. on p. 2).

[ZLZ09] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. "Predicting missing links via local information". In: *The European Physical Journal B* 71.4 (2009), pp. 623–630 (cit. on pp. 1, 2).

Appendix



(a) ROCC Average

(b) PRC Average

Figure 4: ROCC and PRC when predicting future Bitcoin transactions using the mean of all of the graphical features.