Behavior Language Processing with Graph based Feature Generation for Fraud Detection in Online Lending

Wei Min
CreditX
Shanghai, China
minw@creditx.com

Zhengyang Tang
CreditX
Shanghai, China
tangzhy@creditx.com

Min Zhu
CreditX
Shanghai, China
zhumin@creditx.com

Yuxi Dai
CreditX
Shanghai, China
daiyuxi@creditx.com

Yan Wei
CreditX
Shanghai, China
weiyan@creditx.com

Ruinan Zhang
CreditX
Shanghai, China
zhangrn@creditx.com

ABSTRACT

Online lending has exploded in China in recent years. However, the financial agents are vulnerable for fraud attacks which reults in huge financial losses. Anti-fraud detection methods for traditional financial services are less effective against online frauds. As a group effort at CreditX, we designed an accurate, efficient, and scalable online fraud detecting mechnism by delivering a behavior language processing (BLP) framework. Our solution integrates multiple layers from user online behavior data acquisition, knowledge graph building, feature extraction, to final predictive models. As a core component of BLP, we applied graph phomophily theory on selecting social relationships to build a fraud centric bipartite graph. Key graph features are generated by combining graph thoery and experts' domain knowledge to capture linked fraudulous behaviors. The results of online fraud detection on massive real-world data have shown our graph based feature extraction method significantly boosts the accuracy and effectiveness of BLP model.

CCS CONCEPTS

•Computing methodologies → Machine learning approaches; Machine learning algorithms;

KEYWORDS

online lending, financial fraud detection, behavior language processing, graph analysis, homophilic test, feature extraction

ACM Reference format:
Wei Min, Zhengyang Tang, Min Zhu, Yuxi Dai, Yan Wei, and Ruinan Zhang. 2018. Behavior Language Processing with Graph based Feature Generation for Fraud Detection in Online Lending. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web, Marina Del Rey, CA, USA, 2018 (MIS2),* 8 pages.
DOI: 10.475/123_4

MIS2, Marina Del Rey, CA, USA
 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

1 INTRODUCTION

Online lending industry in China experienced a rapid growth in recent years. According to the statistics [1], around 2000 online lending companies made up over 200 billion RMB transactions in the past 10 months of 2017. The tremendous lending helped to provide financial services to the consuming marketing, while on the other hand we also noted this industry is prone to fraudsters' attacks. Since this kind of financial service is for people with no guarantee nor mortgage, and uncovered by traditional credit service, the overdue ratio was estimated over 20% and related loss can be considerable for the lending companies without proper risk control. Meanwhile, traditional models like the logistic regression based scorecard provided by FICO are still heavily adopted in financial institutes as the core risk control strategy, but the solution presents a natural weakness in online lending senario where structured credit data is deficient.

Fraud detection in online lending have many challenges:

(1) **Sparsity in Credit-related Features**: documentations presenting consumers' credit status are usually strong features, such as mortgage, job certification, and social insurance, but those data are usually very sparse for the target population of online lending.
(2) **Velocity, Variety, and Volume of Behavioral Data**: mobile devices and applications nowadays penetrate everyone's daily life. Behavioral data tracked on devices is in a boom at both volume and dimensions. Indeed, behavioral data could be a good candidate to illustrate individual's financial risk since it's reveals the applicants' interests, social relationships, lifestyle and is difficult to forge. In practice, the problem is more about how to integrate these data and apply appropriate data mining methodology to extract financial signals for risk control purposes, since behavioral data like browser behavior logs, sequential location infomation, social logs usually come in complex structures.
(3) **Evolution and Craftiness of Fraudulent Strategy**: increasingly sophisticated fraudsters have developed ways to elude discovery. One scenario includes stealing or purchasing huge amount of mobile numbers, then automating the processes to pass the mobile authentication. In another scenario, fraudsters may fake devices and location identities

using virtual machines. Traditional fraud detection mechanism tends to be less effective in the scenarios mentioned above.

2 BEHAVIOR LANGUAGE PROCESSING

To address challenges under such bussiness circumstances and fraudsters' characteristics mentioned above, we designed the **graph analysis powered behavior language processing** (BLP), a novel, generic, scalable and integrated analysis framework which aims to assist financial institutions to build up a more effective anti-fraud system from scratch based on behavior data, as well as to improve the performance of traditional fraud detection tools. Our proposed BLP framework integrates user online behavior data acquisition system, data integration platform, knowledge graph building, feature extraction, and ensemble predictive model building layers. These components together dive the data thoroughly, involving from components about individual risk extraction to network analytics based linked risk identification , thus form inclusive user credit risk profiles. This framework is capable of providing a industry verified, matured methodology on fraud detection of online loaning, which is rarely achieved by either researchers from academic community or vendors providing credit service solutions.

As illustrated in Figure 1, with the authorization of applicants, a costomized SDK is embedded in the host APP to collect behavior data systematically starting from behavior data acquisition module. These behavior data are processed to a specific data schema and then mapped to financial risk knowledge graph which is predefined according experts' domain knowledge of online lending. The specified knowledge graph serves as a flexible data integration layer. The third module is an automated feature extracting layer, which contains a set of feature adapters designed for integrating experts' domain knowledge and advanced data mining skills in the most efficient way on complex data schema with different data structure. One of the key idea behind BLP framework is to incorporate both advanced individual feature generation methods and group fraud signals into an integrated feature framework. As mentioned above, fraud in China online lending is well-considered and well- organized. Fraudsters' motivations are usually the results of influence from relatives, co-workers and friends. Therefore, fraud detection merely by examining individual features is often insufficient. When traditional analytical techniques fail to detect fraud due to lack of decisive information, social network analysis might give new insights by describing how people are influence by each other. Combining individual features and network features can help improve model fraud predictive performance. This methodology will be discussed in detail in this paper. The top layer of BLP framework is a set of ensemble learning algorithm, which has been proved with high performance, reliability and availability. Our best practice is that boosting tree like GBM [2],LightGBM [3], CatBoost [4] and XGBoost [5] exhibits powerful advantages over traditional logistic regression model on behavior related data.

Broadly speaking, financial fraud detection is never a rare topic in the literatures, and there have been plenty of explorations applying various data mining techniques to enhance the accuracy in predicting fraud risks, such as neural networks [6], bayesian networks [7] and support vector machines [8]. As to graph analytics, several works related with anomaly detection have also been done in recent years, such as the earlier paper about detecting anomalous sub-graphs using variants of the Minimum Description Length (MDL) principle proposed by Noble and Cook [9] and an approach for identifying anomalous nodes upon large graph proposed by Akoglu et al. [10]. The essence of problems these works try to solve exhibits the same characteristics, but the solution to each problem is rather domain-specific and lack the capability to scale widely in utilizing behavior data for fraud detection.

In this paper, we share our in-depth corporation with a leading online financial institution about using graph analysis powered BLP solution for fraud prediction. The online lending product is with max loan amount below $500, and need to be paid back within 14 days. Following the industry standard, applications with payment of overdue at least 5 days are labeled as fraud. The following paper is organized as: in section 3, we will describe how to build graph using behavior linked data, and how to extract graph features in detail; in section 4, the validity of graph analysis as a core component of feature extraction is tested compared with BLP without graph component; and the paper concludes with future studies of BLP and other applications of this framework in section 5.

3 GRAPH ANALYSIS AS BLP FEATURE EXTRACTOR

3.1 Network Building

3.1.1 Graph Relationship Selection. The standard BLP data acquisition system collects a rich behavior data, including but not limited to physical features of mobile device and internet access, social connection logs on mobile device, action traces on host APP, GPS trace of location as well as basic information relates to the applicant. The coverage of these data domain varies given different level of authorization from the applicant. These behavior data provides a rich set of entities, such as applicants' mobile number, applicants' home address, applicants' company address, mobile number of applicants' emergent contacts, device related information such as device id, wifi Mac address, GPS coordinates and so on. These entities are connected by historical application records as well as different social network interactions. Entities and relationships form graph. In the aspect of an unipartite graph model, two applications are connected if and only if they share at least one relation entity, for example, if the home addresses of two applications are the same, there is an edge between these two applications. However, not all of the relations are useful in graph. The reason why graph analys is being so powerful in fraud detection is that fraud exhibits **homophilic effects**, which means fraudsters are generally more socially connected to each other. If graph built by the pre-defined relations exhibits evidence of homophily, the relation is worthwhile to be considered. Mathematically, there are three metrics to measure homophility [14], includes:

- **Homophilic Test**: test whether the observed fraction of cross-labeled edges is significantly less than the expected probability, where cross-labeled edges means two application nodes of edge are with different labels, one is fraud and the other is legitimate.

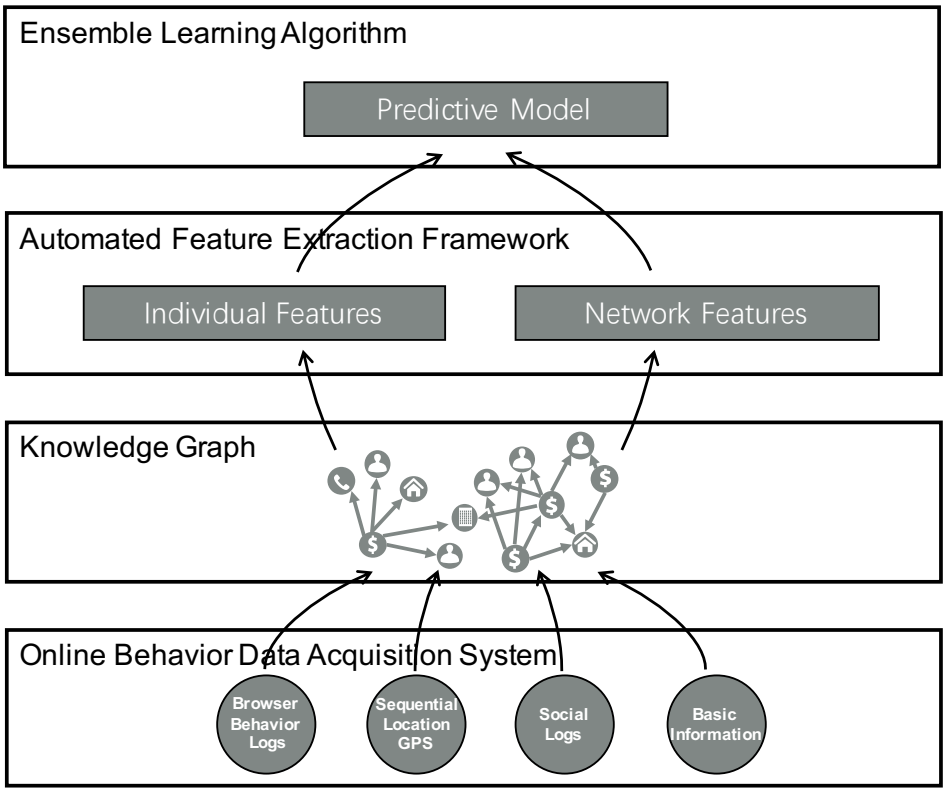


Figure 1: Graph based Behavior Language Processing Structure.

- **Dyadicity**: the observed number of edges with two fraudulent nodes / the expected number of edges with two fraudulent nodes given a random network setting. Dyadicity > 1, indicates that fraudulent nodes are connected more densely among themselves than a random simulation, which exhibits a good characteristic of homophily.
- **Heterophilicity** : the observed number of edges with nodes that have dissimilar label / the expected number of relations that connect applications with dissimilar label given random network. Heterophilicity < 1, indicates that fraud nodes have fewer connections to legitimate nodes than expected randomly, which also approves homophily.

Besides homophily, **connectedness** is another important metric for edge selection. Connectedness measures the density of the network, mathematically, is evaluated as number of observed edges compared with edges from a complete graph configuration. If a pre-defined relationship has low connectedness, then fraud is less likely to spread out. Therefore, useful relationship demonstrates both high homophilic and high connectedness. In reality, high homophilic usually indicates a strong connection with low connectedness. Taking aspect of location coordinate as example, applications are connected if applicants were from the same address during application, when GPS coordinates were collected by data acquisition module. Usually, valid GPS point is with 6 decimal points for both longitude and latitude, and the number of decimal place determines tolerance on accuacy of the address. Different accuracy magnatitude varies in homophily and connectedness. We will compare the following three ways to process GPS data:

- GPS coordinate: an accurate address
- GPS coordinate keep 3 decimal points (GPS_100m): identifies an address with area roughly a 100-meter wide square
- GPS coordinate keep 2 decimal points (GPS_1000m): identifies an address with area roughly a 1000-meter wide square

GPS_100m is favourable beacause both homophily and connectedness are well-considered (see Table 1).

All the pre-defined relationships are selected through homophily and connectedness metrics. Due to company confidential treaty, few typical relationships are illustated in the table (see Table 1). The metrics are calculated on a sampled set with 121164 applications, with 6% application labeled as fraud. The expected fraction of cross-label edge is 0.12. Relationships with observed fraction of cross-label edges smaller than 0.12 and meanwhile with larger connectedness metric are selected. The selection is inline with experts' business awareness, for instance, phone number of company exhibits more homophily than its name because the former is more accurate; wifi Mac address performs better than IP address of the server because the latter is less stable and establishes a relative loose relationship with mobile device.

3.1.2 Bipartite Graph. In above section, relationships are selected in an unipartite graph setting, say graph with only one node type. Two application nodes might be connected by several edges

Table 1: Relation selection by Connectedness and Homophily

Relation	Edge	Connectedness	CrossEdgeFraction	Dyadicity	Heterophilicity
identity number	78964	0.0011%	0.025	13.008	0.208
mobile number	77558	0.0011%	0.023	13.194	0.194
company number	119236	0.0016%	0.086	6.371	0.717
contact number	57304	0.0008%	0.026	13.483	0.216
company name	365786	0.0050%	0.104	2.702	0.867
company address	71886	0.0010%	0.072	6.436	0.597
device id	79408	0.0011%	0.028	12.698	0.233
ip address	168894	0.0023%	0.089	3.145	0.742
wifi MAC	33368	0.0005%	0.035	13.314	0.289
GPS	8378	0.0001%	0.018	16.139	0.151
GPS_100m	39172	0.0005%	0.046	11.650	0.379
GPS_1000m	370186	0.0050%	0.093	2.347	0.774

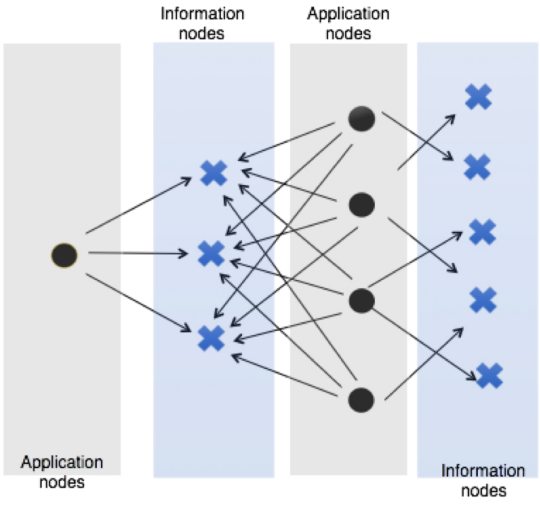


Figure 2: A Bipartite Graph is built by two types of nodes, nodes of the same type can only be connected through nodes with different type, in BLP solution, there are application nodes and information nodes.

when sharing several common relationships. Risk leads mining and exploration becomes abscure in this type of graph model setting. A complex graph is a more suitable graph model in the context. In complex graph, all relation entities such as device id, wifi mac address are also nodes, while application nodes cannot be connected to each other directly, they must be connected through an relation entity. These relation entities are treated as a same node type, say, information node. Then the complex graph is simplified to a bipartite graph. In the graph, application nodes have attributes like application datetime, loan decision(approve/reject), loan performance (fraud/legistimate), loan amount and so on. Attributes of information nodes varies due to the difference of entities. Edge from application node to information node by nature indicates its relationship type. A direct bipartite graph (See Figure 2) setting provides not only better visualization but also solid graph theoretical foundation in fraud detection application.

3.1.3 Edge Weight Setting. The weight of edge in the bipartite graph represents the intensity of the relations. The intensity reveals two characteristics, one is the connection strength of the relation, to explain, relationships with identity number are closer than relationships with company name. In fraud detection setting, the closeness of relationship is estimated by a mapping function from homophilic metrics. Another characteristic to consider is the time decay effect. Fraud is time - dynamic, historical information of the network should be decayed or reweighted based on its recency. The following exponential function is used to esitimate edge weights of the dynamic network:

$$w = a \times e^{-b}$$

where a is the closeness of relation estimated through homophilic metrics, and b is time decay coefficient.

3.1.4 Hubs Removal. In graph theory, degree of nodes follows power law, it stays valid in bipartite networks setting. The degree of an information node summarizes the number of application nodes connected to the information node, which also follows a power law. Taken company name as example, companies with large scale such as top insurance companies and leading logistic companies usually associate with massive loan applications (see Figure 3). Propagation algorithm for fraud spreading globally is degree dependent, nodes with large degree spread proportionally more fraud than low degree nodes. Therefore, hubs of information node will be exposed to large fraction of fraud which arises false alarm. Head-tail break [11] algorithm is an effective algorithm to bin pareto distribution, it helps to detect big hubs automatically. Hubs of information nodes are removed from graph building.

3.2 Graph Fraud Feature Extraction

In this section, we will discuss how to extract fraud risk related features of application node based on network-based analysis. There are mainly three kinds of techniques:

- **Local Metrics**: measures the characteristics of n-order neighborhood around the application node. Given the ego-network of the application node, there are many graph metrics to evaluate the local network structure, such as degree, quadrangle, density. Features are extracted from three

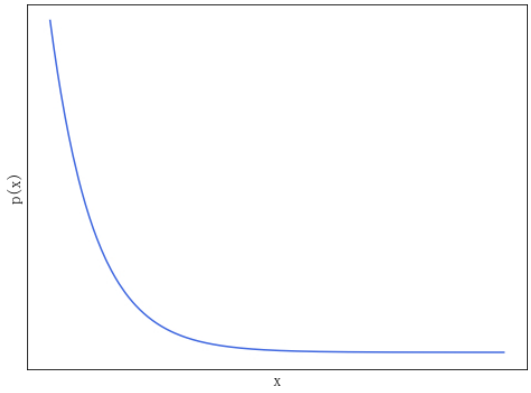


Figure 3: Degree of Company name Node Distribution: the degree of Bipartite Graph also follows Power law.

different angles: basic statistics, label dependency[13], and edge strengthness weighted, see the feature sets in Table 2.

- **Degree Related Features**: In bipartite graph, the first-order neighbors are information nodes, neighbor size measures the distinct information associated with the application; the second - order neighbors are applications that are shared same same information with the target application node.
- **Quadrangles** : A quadrangle in bipartite is a subgraph with two application nodes connected by two different information nodes. Quadrangle investigates the connection strength between two applications.
- **Local Cluster Coefficient** : Another neighborhood metric to evaluate the network's local density is called cluster coefficient. The density mertic is calculated as the observed connectedness of the subgraph compared with the expected connection in a complete graph setting.

- **Global Metrics**: Given a network with historical labeled fraudulent application nodes, how can we use this knowledge to infer a primary fraud probability for the unlabeled application nodes. Personalized page rank algorithm is used to spread fraud from the labeled fraud application nodes to information nodes, and then to unlabeled application nodes propotional to the relationship strength while simultaneously decaying the weights of past frauds. The primary fraud probability exposed to the unlabeled nodes is called fraud score. This metric was proven effective in Gotach framework [12].
- **Mis-match Defined By Human Expertise**: In risk management, finding leads for mis-match is an effective way to detect fraud. There are two aspects of mis-match. One is caused by information collected from different channels. Jaccard distance is used to mathematically quantify the similarity of a given type of information from different data sources (similarity of two sub-graph). Another way mis-match can be caused is that individual information conflicts with the rest of network.

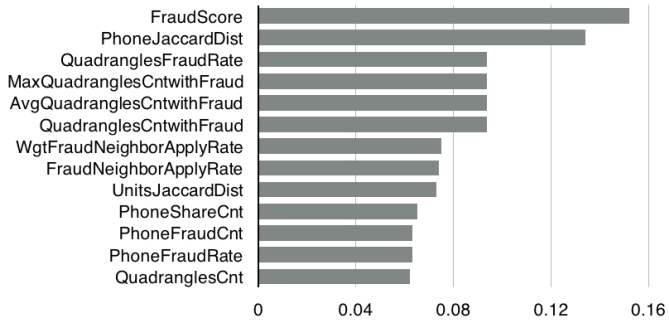


Figure 4: Example of Graph features with top information value.

Totally, hundreds of graph features are extracted from the bipartite graph, information value (IV) is used to evaluate feature effectiveness. Top features are illustrated in Table 4. *FraudScore* measures how the application affects by fraud from the rest of networks ranks top, high fraudscore is a strong indicator of linked fraud. Followed by *PhoneJaccardDist*, which measures how identity of phone number related information filled by the applicant on the application form to the same information collected from credit bureau. *PhoneJaccardDist* is one kind of mis-match metric defined by expert. It is particularly important to bridge the richness of experts'knowledge to the technical limitations of network analytics by selecting the most relevant data features for the analysis.

4 FRAUD PREDICTION MODEL RESULT

In this section, parts of the result that adopting BLP as fraud detection solution are shared. Nine months historical applications with matured loan performance were extracted (roughly **13.5 million applications**), which resulted in a bipartite graph with **100 million nodes** , and **150 million edges**. Applications from the 7th to the 9th month are sampled for fraud prediction modeling (75% for training, 25% for testing). Due to the fact that network is dynamic, graph features for each application in training data are extracted based on its own past 6 months graph snapshot, so that each application has the same observation time window. Individual features are then extracted automatically using feature adaptors from BLP feature extraction module. Before adopting BLP solution, fraud detection in this financial institution heavily relied on experts' experience, with no effective and systematic methodology to control online fraud attack. Other financial agents are also powerless to process behavior data into valid risk signal even though behavior data are collected. Therefore, there is no industry standard to be compared with BLP results. Since graph analysis is introduced as core component of BLP, the experimentation is set to illustrate why graph analysis is capable of empowering BLP for fraud detection.

- BLP_base: ensemble model built on BLP individual feature components.
- BLP_graph: ensemble model built on integrated BLP feature layer with both individual and graph features.

Both models are trained with the same ensemble model framework, LightGBM, the state of art classifier from BLP model module.

Table 2: Graph Local Feature Extractions

Metrics	Degree	Quadrangles	Local Cluster Coefficient
basic statistics	– number of associated applications	– total number of quadrangles – the max/mean/average of quadrangles frequency with the associated application nodes	– 2-order neighbors cluster coefficient – 3-order neighbors cluster coefficient
label dependent	– fraction of associated fraud applications	– fraction of quadrangles associated with fraud application nodes – the max/mean/average quadrangles frequency with the associated fraud application nodes	– 2-order neighbors (with only information nodes and fraud application nodes)cluster coefficient – 3-order neighbors (with only information nodes and fraud application nodes)cluster coefficient
intensity weighted	– weighted number of associated applications – weighted fraction of associated fraud applications	– weighted total number of quadrangles – the weighted max/mean/average quadrangles frequency of the associated application nodes – the weighted max/mean/average quadrangles frequency of the associated fraud application nodes	

4.1 Results

4.1.1 Model Performance. The result is shared in Table 3. Considering AUC as the performance metric, BLP_graph model is 6% better than BLP solution without graph features (see Figure 5), and **Max KS** (mertic derived from kolmogorov - smirnov Test) as financial industry standard metric is improved by **27%** , which means that the BLP_graph model significantly boosts fraud predictive ability. Feature importance evaluated using information gain from boosting tree helps peek the black-box model to understand feature contribution, see feature importance list in Figure 6. Graph features that rank top are: *FraudScore* that measures fraud obsorbed from the whole network through propagation algorithm, *PhoneJaccardDist* means the consistency of phone information collected from different data sources, *QuadranglesFraudRate* quantifies the connection closeness of the target application node with historical fraud application nodes. These features are strong signals for linked risk. Other individual features like *connection_* generated by BLP feature adaptors are not in the scope of the paper.

4.1.2 Model Stability. Besides fraud predictive ability, stability is also a key factor of an effective fraud detection mechnism. There are multiple ways to evaluate model stability.

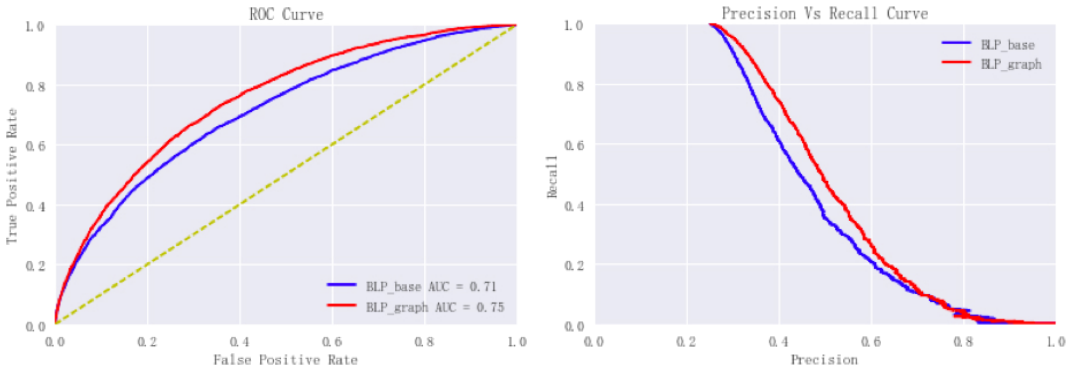
(1) **Preditive ability in out-of-time window dataset**. Subsequent 6 month histrotical applications were inserted to graph database. Applications from the 13th to 15th month were sampled as an out-of-time window held-out set (3 month gap with data for modeling). Max KS of BLP_graph is dropped by 16% from testing set to held-out set, compared with a much larger drop (23%) for BLP_base, which indicates **graph features are more robust** than individual featuers. To notice that, the decay of BLP models preditive ability is acceptable for online lending scenario given the rapidly changed market.

(2) **Feature stability** is critical for model stability. In financial risk modeling, Population Stability Index (PSI) is usually used to evaluate feature distribution drift. In this experimentation, feature PSI are calculated on its' distribution from training set and held-out set. All of the graph feature PSI are less than 0.05, which indicates the stability of graph features.

(3) **Model transfer ability** is also a key metric in terms of model stability. Both BLP_base and BLP_graph models in the experiment were applied on another online lending product. The two products are similar except targeting to different geographic channels. The transfer ability evaluated by Max KS of the second lending product proves the robustness of BLP_graph again.

5 CONCLUSION

In this paper, a sophisticated behavior language process framework which integrated graph analysis was introduced to solve online lending fraud attack. We started by explicating the challenges of fraud detection in online lending scenario: with limited credit data, financial agent is extremely vulnerable to fraud attack. Traditional strategy is insufficient in both getting valid data and adopting systematic methodolog to capture the emerging pre-planned and well-organized online lending fraud attack. With BLP, a framework integrates behavior data collection, data integration, feature extraction and model building to handle unstructured behavior data for online fraud detection. Graph analysis as an effective method to



(a) AUC Curve Comparison **(b) Precision and Recall Curve Comparison**

Figure 5: Model Performance Comparison: BLP_graph significantly boosts performance of individual feature based preditive model

Table 3: Experimentation Results

Models	Features	Test AUC	Test KS	Held-out KS
BLP_base	1774	71%	0.30	0.23
BLP_graph	1852	75%	0.38	0.32

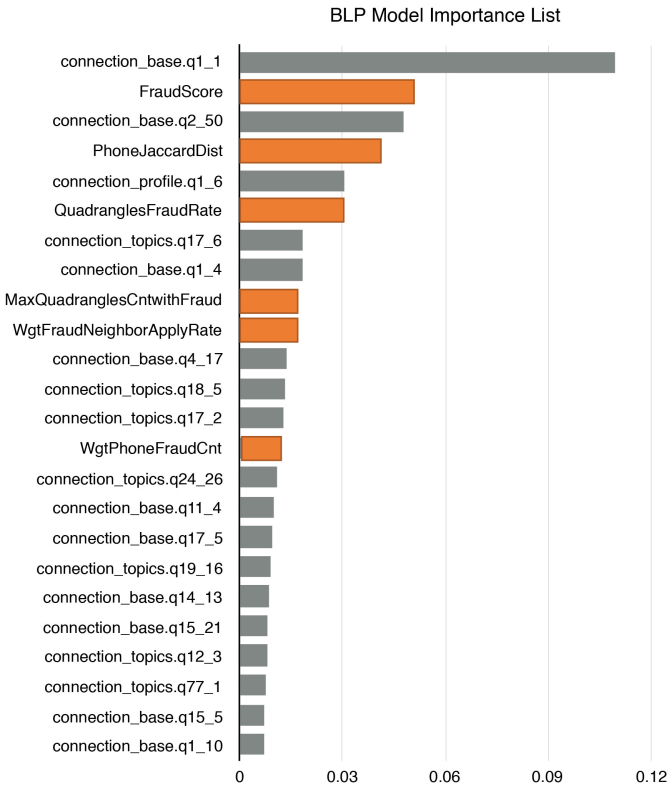


Figure 6: Feature Comtribution List (top features): dark orange for graph features v.s. grey for individual features

process behavior data was highlighted on 1) design and select effective relations for network building; 2) set edge weight to represent link strength and caputure time recency effect; 3) extract graph features by integrating graph theory and experts experience. We demonstrated the improvement of graph analysis as feature extraction for fraud predictive model in production level data, it boosts the main metric of interest (Max KS) by 27% compared with BLP without graph features. One of future work is to revise *FraudScore* by using degree independent personalized page rank algorithm [12] for fraud propagation to reduce false alarm. Another future work focuses on further graph feature extraction, such as node embedding[15], which recently is emerged as a powerful representation of graph-structure data for supervised learning tasks. Though the BLP framework is customized for online lending fraud detection senario, it can easily be migrated to other online lending senario such as credit monitoring and marketing, and other online fraud detection senario such as e-Bussiness.

REFERENCES

[1] Tencent Techonology.The boom of online lending in China.http://new.qq.com/omn/20171130A0R1K1.html
[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In Advances in Neural Information Processing Systems (NIPS), pp. 3149-3157. 2017.
[3] Jerome H Friedman. " Stochastic gradient boosting. " Computational Statistics & Data Analysis, 38(4):367-378, 2002.
[4] Dorogush, Anna Veronika, et al. " Fighting biases with dynamic boosting." arXiv preprint arXiv:1706.09516 (2017).
[5] Tianqi Chen and Carlos Guestrin. " XGBoost: A Scalable Tree Boosting System." In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016
[6] Ghosh, and Reilly. 1994."Credit Card Fraud Detection with a Neural-Network" In 1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, 3:621-30.

[7] Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. 2007. "Data Mining Techniques for the Detection of Fraudulent Financial Statements." Expert Systems With Applications 32 (4): 995-1003.
[8] Chen, Rong-Chang, Tung-Shou Chen, and Chih-Chiang Lin. 2006."A NEW BINARY SUPPORT VECTOR SYSTEM FOR INCREASING DETECTION RATE OF CREDIT CARD FRAUD." International Journal of Pattern Recognition and Artificial Intelligence 20 (2): 227-39.
[9] Noble, Caleb C., and Diane J. Cook. 2003. "Graph-Based Anomaly Detection." In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 631-36.
[10] Akoglu, Leman, Mary McGlohon, and Christos Faloutsos. 2010. "OddBall: Spotting Anomalies in Weighted Graphs." In PAKDD'10 Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II, 6119:410-21.
[11] Bin Jiang. 2010. "Head/tail Breaks: A New Classification Scheme for Data with a Heavy-tailed Distribution".
[12] Vlasselaer, Vronique Van, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. 2017. "GOTCHA! Network-Based Fraud Detection for Social Security Fraud." Management Science 63 (9): 3090-3110.
[13] Kajdanowicz, Tomasz, Przemyslaw Kazienko, and Piotr Doskocz. 2010. "Label-Dependent Feature Extraction in Social Networks for Node Classification." Social Informatics, 89-102.
[14] Baesens, Bart, Vronique Van Vlasselaer, and Wouter Verbeke. 2015. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Vol. 0.
[15] Thuy Vu, D. Stott Parker. 2015. "Node Embeddings in Social Network Analysis." ASONAM '15, ISBN 978-1-4503-3854-7/15/08