Comparison of anomaly detection performance based on edge weight

Jeong Ii tae

GraphUserGroup

Contents

1. Introduction

Anomaly detection 분야에서 왜 그래프가 효율적일까?

2. Related research & Preliminary

2.1.Graph based anomaly detection overview

2.2.Random-walk based method vs. graph neural network based method

2.3.node2vec with abnormal detection

3. Methodology

3.1.node2vec, node2vec+ (numba)

3.2.Random walk handling (biased)

3.3.Edge design

4. Experiment

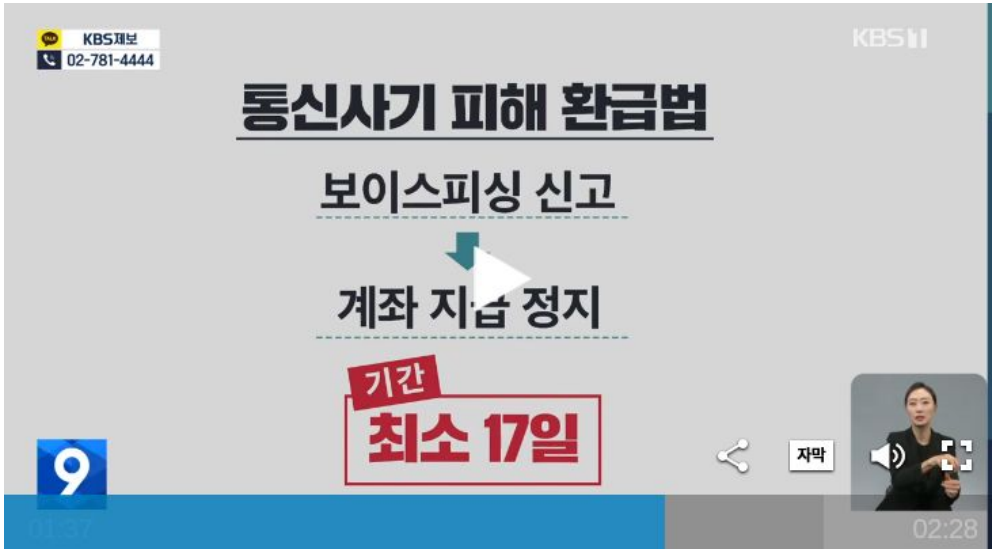
...etc

5. Conclusion

Abnormal transactions capture always challenging problems in financial industry

[제보K] 난데없는 입금 뒤 '묻지마' 계좌정지…금융위 "대책 마련할 것"

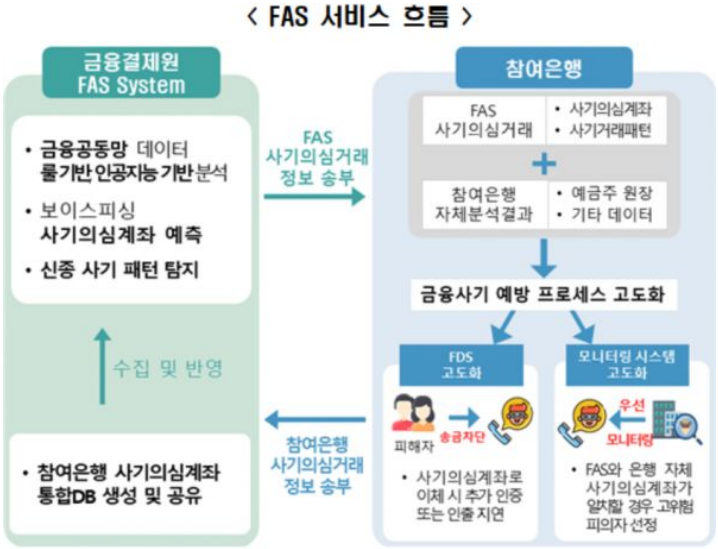
입력 2023.01.09 (21:43) | 수정 2023.01.19 (11:03) 뉴스 9



https://news.kbs.co.kr/news/view.do?ncd=7138185

금결원, 13개 은행에 보이스피싱 근절 위한 서비스 제공

이진희 기자 | ljh@seoulfn.com | 승인 2021.02.15 09:54 | 댓글 0



https://www.seoulfn.com/news/articleView.html?idxno=410629

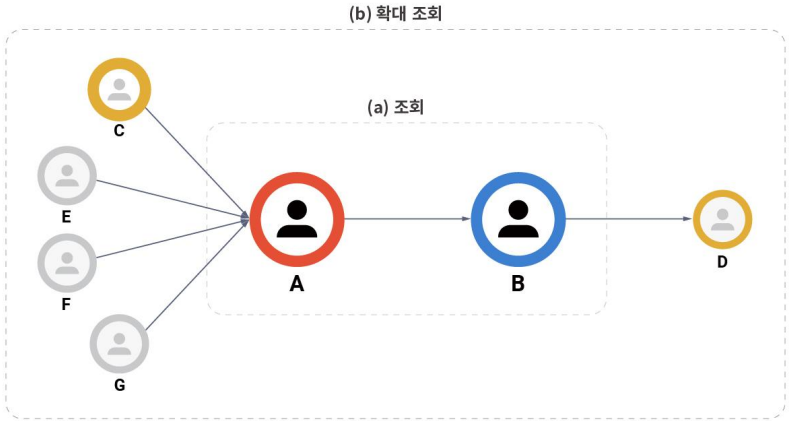
Anomaly detection 분야에서 왜 그래프가 효율적일까? Why graph?

→ 거래들간의 관계를 파악하는데 최적의 자료구조이기때문.

정형 데이터 관점

scanning cost : R^{n,k}

비정형 데이터 관점



scanning cost : R^{edge}

‘관련’있는 유저들만을 탐색하여 정보를 빠르게 분별한 후 판단하기에 용이함.

1.'빠르게' 이상 거래를 탐지하고 액션을 취해야한다.

2. 이상거래의 기준인 '룰' 의 트렌드가 반영되어 모니터링 시스템에 이상거래 추세가 반영이 되어야함.

→ 데이터 관리와 기계학습 방식이 잘 조합되어야 한다.

이상거래 탐지 기계학습 메뉴얼

1.데이터 수집: 이상을 탐지할 데이터셋을 수집합니다.

2.데이터 준비: 데이터를 정리하고 정리하여 분석에 적합한 형식으로 만듭니다.

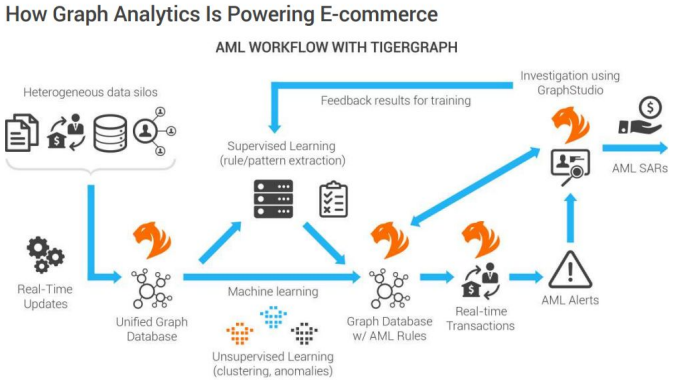
3.정상 동작 식별: 정상 동작을 포함하는 데이터셋의 일부분을 사용하여 이상 탐지 모델을 학습시킵니다. 모델은 어떤 것이 일반적이고 예상되는 동작인지를 학습합니다.

4.이상 탐지: 학습된 모델을 나머지 데이터셋에 적용합니다. 모델은 각 데이터 포인트를 학습한 정상 동작과 비교하여 크게 다른 데이터를 식별합니다.

5.이상 분석: 탐지된 이상을 분석하여 이상의 특성과 가능한 원인을 파악합니다.

6.대응: 분석 결과에 따라 적절한 조치를 취하여 이상을 해결하고 부정적인 영향을 최소화합니다.

Graph ML + DB process



그래프 관점으로 이상 거래탐지 직접해보자.
시작전, 다른곳에서는 어떻게 하고 있는지
살펴보자. 잘 될까?

node, edge , sub-graph , graph 총 4가지 관점에서 접근할 수 있음.

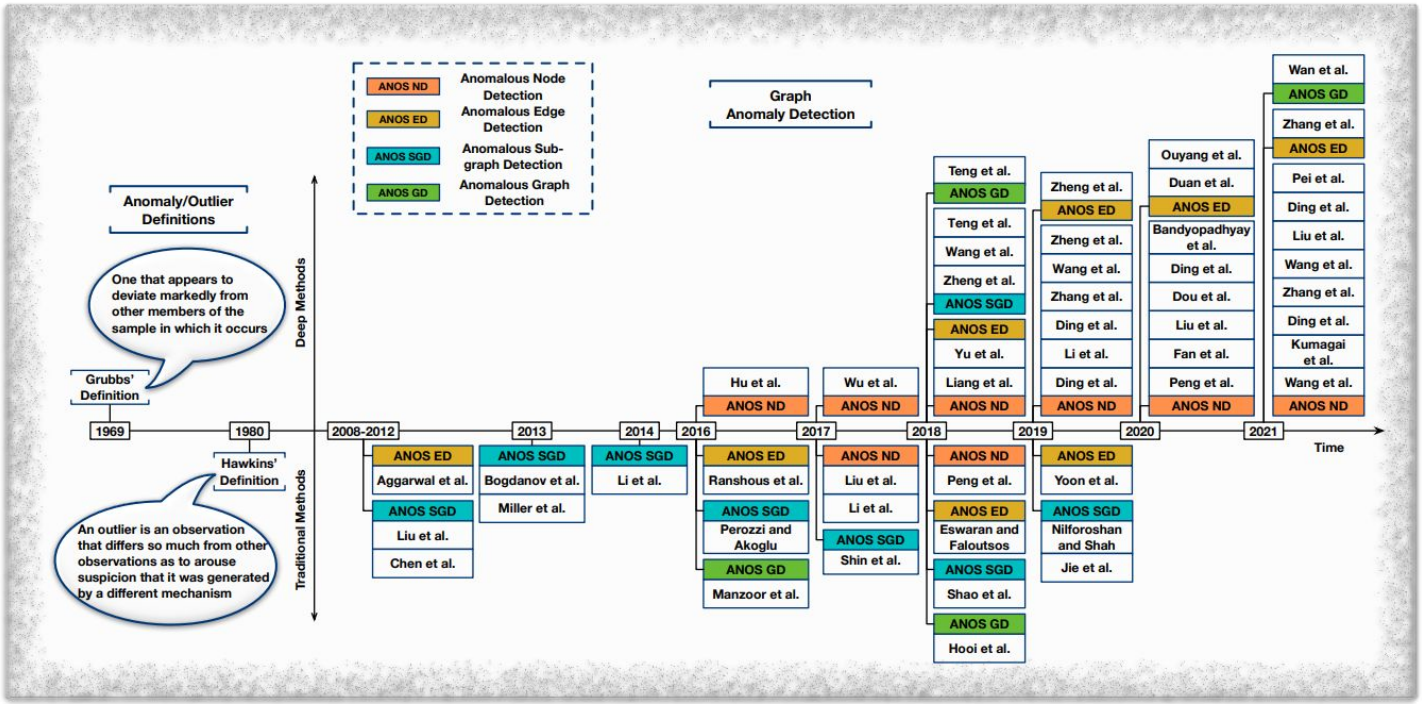
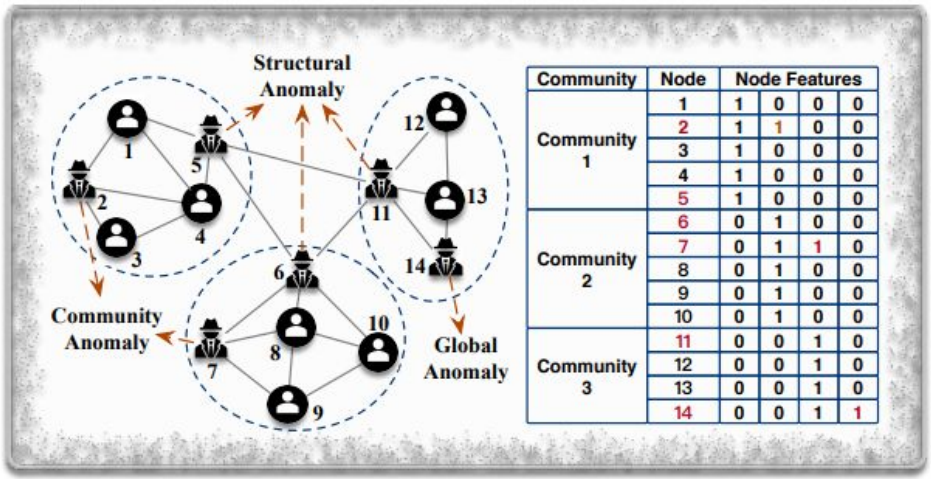


Fig. 2: A Timeline of Graph Anomaly Detection and Reviewed Techniques.

* Ma, Xiaoxiao, et al. "A comprehensive survey on graph anomaly detection with deep learning." IEEE Transactions on Knowledge and Data Engineering (2021).



Community	Node	Node Features			
Community 1	1	1	0	0	0
	2	1	1	0	0
	3	1	0	0	0
	4	1	0	0	0
	5	1	0	0	0
Community 2	6	0	1	0	0
	7	0	1	1	0
	8	0	1	0	0
	9	0	1	0	0
	10	0	1	0	0
Community 3	11	0	0	1	0
	12	0	0	1	0
	13	0	0	1	0
	14	0	0	1	1

- Global anomalies only consider ***the node attributes.*** They are nodes that have attributes significantly different from all other nodes in the graph.

- Structural anomalies only consider ***the graph structural information.*** They are abnormal nodes that have different connection patterns (e.g., connecting different communities, forming dense links with others).

- Community anomalies consider ***both node attributes and graph structural information***. They are defined as nodes that have different attribute values compared to other nodes in the same community.

* Ma, Xiaoxiao, et al. "A comprehensive survey on graph anomaly detection with deep learning." IEEE Transactions on Knowledge and Data Engineering (2021).

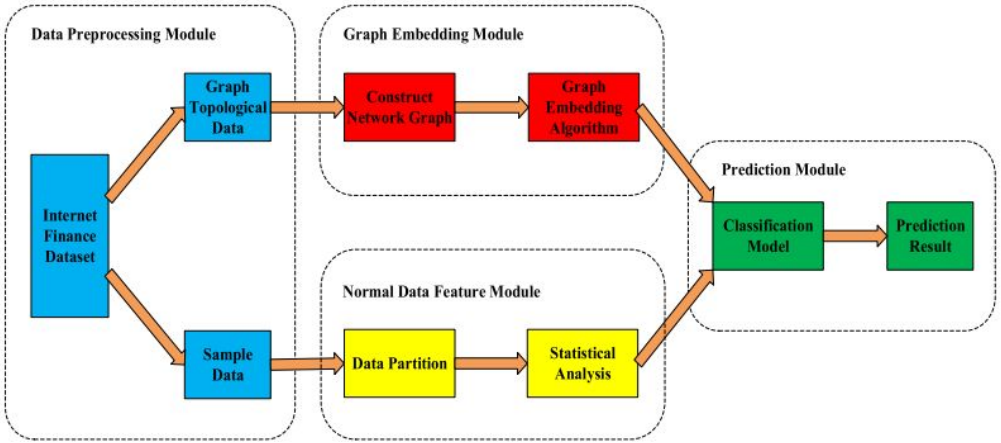
structural information 만을 사용할지, attribute 를 추가적으로 활용할지의 차이로 나뉨.

Random-walk

Graph Neural Network

가용할 수 있는 자원은 오직 cpu…!

Random-walk based method



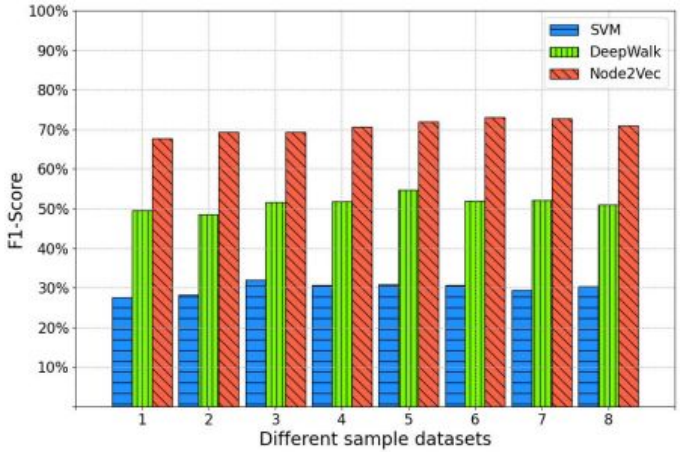


FIGURE 7. Evaluation on F1-score with different datasets.

Node2Vec algorithm on Spark GraphX to learn and represent the topological features of a vertex in the network graph into a low-dimensional dense vector.

* H. Zhou et al.: Internet Financial Fraud Detection Based on Distributed Big Data Approach With Node2vec

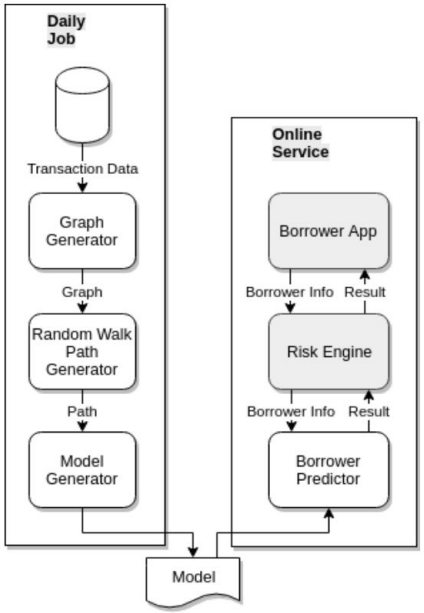


Figure 1. System architecture

Figure 4. Attributes organization.

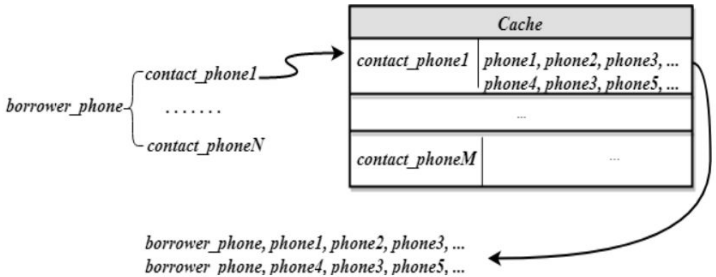


Figure 5. Random walk simulation

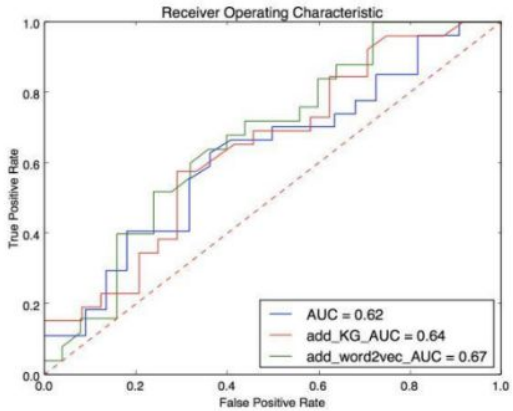


Figure 9. Improvement on AUC.

The original Node2Vec path only contains vertex ID, meaning only graph structure is taken into account. Actually, vertex attributes, such as third-party credit score, sex, etc, can all be important. **We will combine attributes to a long value, as Figure 4 shows. If the raw attribute has many values such as credit score, it will be divided into several levels.**

→ graph structure design (node attributes,property 를 각각의 노드로 빼서 random-walk 에 적용함.)

* Zhan, Qing, and Hang Yin. "A loan application fraud detection method based on knowledge graph and neural network." Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence. 2018.

Table 3: Macro-F1 on sampled dataset with sampled training ratios.

Algorithms	20%	40%	60%	80%
node feature skip-gram	0.628	0.663	0.668	0.67
matrix factorization	0.518	0.57	0.614	0.617
LINE	0.737	0.74	0.742	0.745
node2vec	0.74	0.741	0.743	0.745
FeatNet-1	0.742	0.775	0.801	0.867
FeatNet-2	0.745	0.787	0.829	0.875

(5) FeatNet-1. FeatNet-1 uses device ID and Wi-Fi AP relationships.
(6) FeatNet-2. FeatNet-2 uses device ID, Wi-Fi AP and user account relationships.

→ graph structure design
(user-user 에서 user property 를 노드, 엣지로 빼냄.)

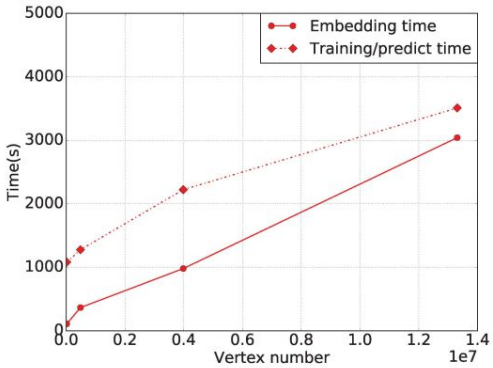


Figure 2: Scalability.

Table 5: AUC values on large-scale dataset with sampled training ratios.

Algorithms	20%	40%	60%	80%
node feature skip-gram	0.698	0.700	0.699	0.705
FeatNet-1	0.874	0.878	0.880	0.881
FeatNet-2	0.875	0.880	0.883	0.885

Some optimization from previous work [15] makes the sampling procedure efficient. As shown in Figure 2 FeatNet-2 scales nicely with learning time increasing linearly, and completes learning of the whole dataset in 59 minutes.

* 13,567,732 device IDs, 1,260,835 of which are spam; 12,306,897 of which are not spam. There are 31,427,140 BSSIDs and 136,006 user account IDs.

* Xu, Chao, et al. "FeatNet: large-scale fraud device detection by network representation learning with rich features." Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security. 2018.

결국 structural information 을 잘 반영하기 위해선, graph design 이 중요하다.

선택한 방법론
1. random-walk based 를 잘 활용하기 위한 묘책이 없을까 ?
2. DB 와 연동(scalability)할 수 있는 방법론이 없을까?
3. graph-design handling 의 중요성(상황에 따라 파라미터를 유동적으로 활용할 수 있어야함.)을 담고 싶다.

→node2vec+ with Edge weight.

biased random walk , p와 q를 통해 node representation learning 을 하는 알고리즘.
본 과정을 통해 어느 노드로 이동(전이)할지에 대한 확률이 계산되어 이동함.

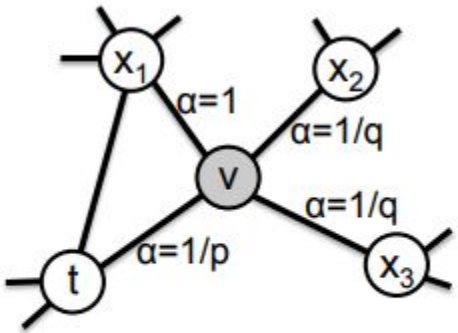


Figure 2: Illustration of the random walk procedure in *node2vec*. The walk just transitioned from t to v and is now evaluating its next step out of node v. Edge labels indicate search biases α.

Return parameter, p. Parameter p controls the likelihood of immediately revisiting a node in the walk.

Setting it to a high value (> max(q, 1)) ensures that we are less likely to sample an already visited node in the following two steps (unless the next node in the walk had no other neighbor).

This strategy encourages moderate exploration and avoids 2-hop redundancy in sampling. On the other hand, if *p is low (< min(q, 1)), it would lead the walk to backtrack a step (Figure 2) and this would keep the walk "local" close to the starting node u.*

In-out parameter, q. Parameter q allows the search to differentiate between "inward" and "outward" nodes.

Going back to Figure 2, *if q > 1, the random walk is biased towards nodes close to node t. Such walks obtain a local view of the underlying graph with respect to the start node in the walk and approximate BFS behavior* in the sense that our samples comprise of nodes within a small locality.

In contrast, if *q < 1, the walk is more inclined to visit nodes which are further away from the node t. Such behavior is reflective of DFS which encourages outward exploration.*

* Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

PecanPy: A parallelized, efficient, and accelerated *node2vec(+)* in Python

Learning low-dimensional representations (embeddings) of nodes in large graphs is key to applying machine learning on massive biological networks. *Node2vec* is the most widely used method for node embedding. PecanPy is a fast, parallelized, memory efficient, and cache optimized Python implementation of *node2vec*. It uses cache-optimized compact graph data structures and precomputing/parallelization to result in fast, high-quality node embeddings for biological networks of all sizes and densities. Detailed source code documentation can be found here.

The details of implementation and the optimizations, along with benchmarks, are described in the application note *PecanPy: a fast, efficient and parallelized Python implementation of node2vec*, which is published in *Bioinformatics*. The benchmarking results presented in the preprint can be reproduced using the test scripts provided in the companion benchmarks repo.

v2 update: PecanPy is now equipped with *node2vec+*, which is a natural extension of *node2vec* and handles weighted graph more effectively. For more information, see *Accurately Modeling Biased Random Walks on Weighted Graphs Using Node2vec+*. The datasets and test scripts for reproducing the presented results are available in the node2vec+ benchmarks repo.

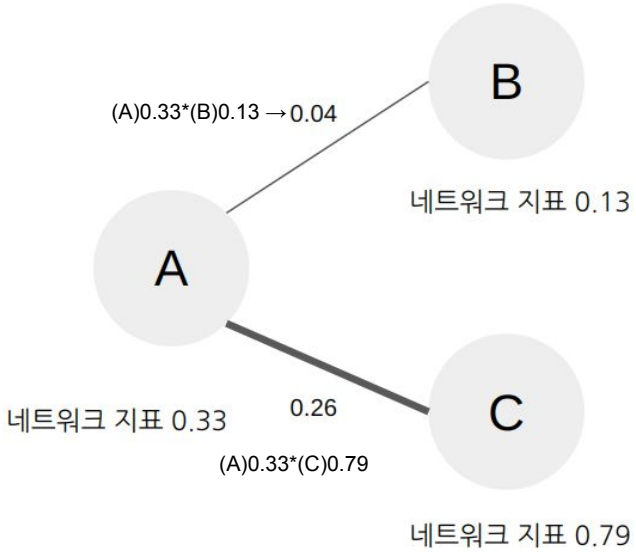
random-walk 시 노드간 이동할 확률, 전이확률(transition probability) 계산을 병렬처리(parallelized)할 수 있는 알고리즘. (all numba)

pre-computation cost 가 개선된 버전. (node2vec+)

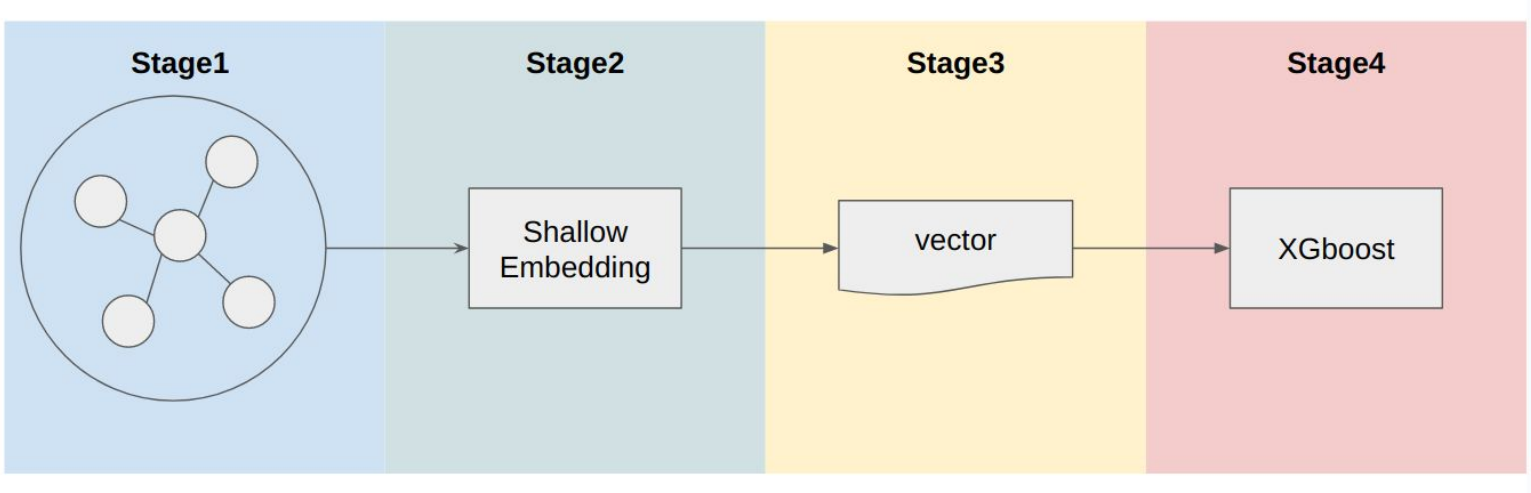
기존 알고리즘(node2vec) 에서 cost 비중이 높았던 부분이였음.

* Liu, Renming, Matthew Hirn, and Arjun Krishnan. "Accurately modeling biased random walks on weighted networks using node2vec+." Bioinformatics 39.1 (2023): btad047.

노드간 전이확률을 튜닝하기 위해, edge weight design 을 진행함. 각 노드의 centrality 를 활용하여 상호 관계가 있는 노드들끼리 내적. 도출된 값을 edge weight 로 부여함.



예) 노드 A 와 연결된 B,C 중
산술기하평균 값을 엣지로 넣어주었을 때,
A-C 가 0.26으로 크기에 랜덤워크시,
A에서 C로 갈 확률이 더 크다.



Stage1: 주어진 그래프 데이터에서 임베딩 수행시 사용할 가중치 특성을 추출. 가중치 특성은 고유 벡터 중심성을 의미.

Stage2: 얕은 임베딩(Shallow embedding) 를 진행. 가중치를 포함한 모델인 Node2vec+와 가중치를 포함하지 않은 모델 Node2vec을 각각 활용, 임베딩 값 추출. * 가중치는 Stage1에서 추출한 고유 벡터의 중심성으로, 이를 엣지 특성에 반영하여 임베딩을 수행한다.

Stage3: 추출된 임베딩값을 저장하고 전처리 단계.

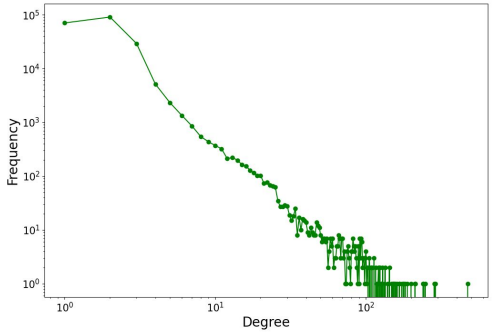
Stage4: 임베딩 값을 XGboost 분류기에 적용하여 분류 결과와 정답지를 대조하는 단계.

ELLIPTIC

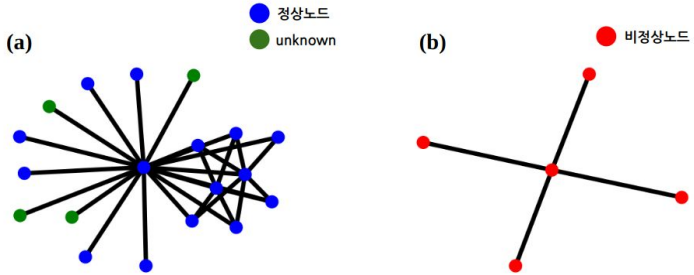
노드 종류	갯수	비율(%)
정상	42,019	21%
비정상	4,545	2%
unknown	157,205	77%

There are 203,769 node transactions and 234,355 directed edge payments flows.

- updated version :
https://github.com/git-disl/EllipticPlusPlus/tree/main



power-law (멱함수) 를 띔. 소수의 노드가 네트워크 내에서 정보 허브 역할을 하여 많은 영향력이 있다고 볼 수 있다

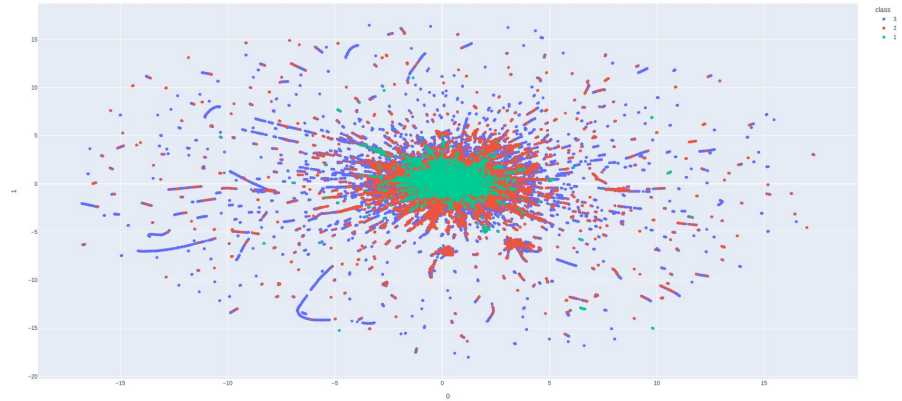


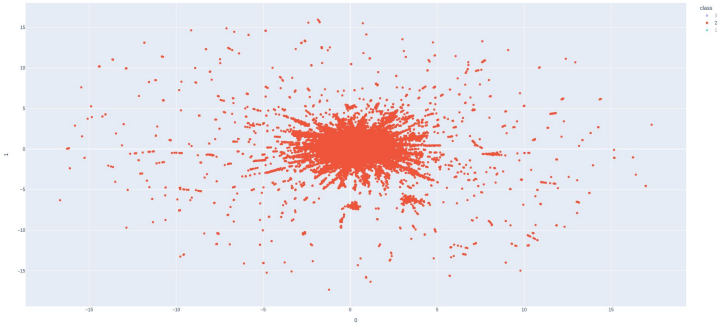
(a) 정상 노드와 unknown의 거래가 발생한 사례. 정상 노드끼리 군집을 이루며 거래가 이루어졌다.

(b) 비정상노드들의 거래가 발생한 사례. 비정상노드들은 비정상노드들끼리 거래가 이루어짐을 관찰. → 비슷한 특성을 지닌 노드들끼리 거래가 발생함을 유추.

* M. Weber et. al. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics, KDD 19 Workshop on Anomaly Detection in Finance, 2019.

T-SNE 2차원 시각화 결과, 몇몇 노드들을 제외하고는 응집이 잘 되있음을 간접적으로 확인할 수 있음.

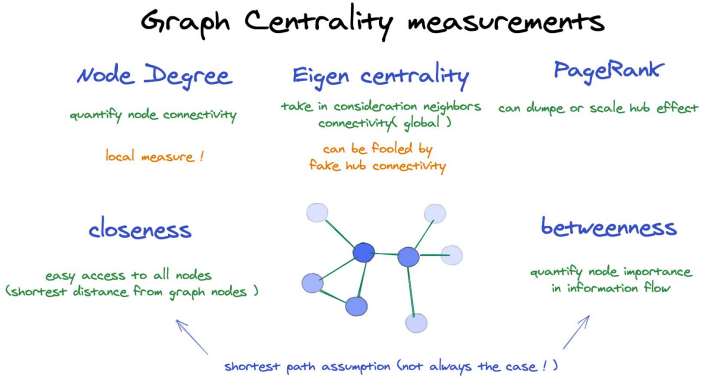




정상

비정상

centrality 를 반영한 결과 중 hop 의 distance 를 반영한 closeness 가 성능 감소측면에서 선방함.



*https://towardsdatascience.com/notes-on-graph-theory-centrality-measurements-e37d2e49550a

다양한 centrality 를 반영했을때. (unofficial)

성능지표	raw-feature	eigenvector	betweenness	degree	closeness	raw-node2vec
roc-auc	0.9453	0.9190	0.9135	**0.9460**	0.9446	0.9436
f1-score	0.9581	0.9680	0.9651	0.9449	0.8685	**0.9728**
undersampling (roc-auc)	0.9337	0.8343	0.9071	0.9343	**0.9357**	0.9341

cugraph vs. **networkx with ray**

networkx + ray 를 조합한다면, GPU 가 없어도 large-graph 에서 centrality 를 추출하는데 무리가 없음.

eigenvector 를 edge weight 로 주었을 때. (학술대회용)

	Node2vec	Node2vec+
f1-score	0.9729	0.9556
auc-roc	0.9453	0.8763

** real-world 데이터로 실효성을 검증해보고싶습니다.. 데이터셋 지원이 가능하신분들.. 재밌는결과 가져다드리겠습니다(free...!)

Edge weight (graph structure) 에 따라 결과값이 달라지므로, graph design 은 중요한 요소이다.

이상거래 탐지 추세는 Mixer (다수의 대포통장을 경유하여, 최종 종착 통장인 현금으로 전환할 통장으로 도착하는 추세임).

즉, hop 에 특화된 1.distribution capture 과 2. random-walk biased optimization 가 조합된다면 더욱 좋은 결과를 추출할 수 있을것으로 기대됨.

closeness 가 성능측면에서 타 centrality 대비 높은걸로 보아 유의미한 상관성이 있을것이라고 볼 수 있음.

(rewiring technology etc… 본 주제를 가지고 같이 후속 연구하실분을 구합니다.)

GUG seminar 2회차는 9월 중순에 진행될 예정입니다. 많은 관심 부탁드립니다.
연사 및 포스터 모집!!!!!!중입니다!!!!!

오픈소스커뮤니티 <comm@oss.kr>
나에게

안녕하세요, 정이태님.
Open UP 커뮤니티 담당자입니다.

현재 OSS 포털사이트에 커뮤니티 등록이 완료되었습니다.

2023년도 공개SW 커뮤니티 하반기 지원 안내 드리겠습니다. 😃

[2023 SW 커뮤니티] GraphUserGroup

통합지원센터 관리자 | 2023-06-21 19:05:06 | 17

커뮤니티명	GraphUserGroup
홈페이지 및 SNS주소	https://www.graphusergroup.com/
프로젝트 저장소	https://github.com/graphusergroup
커뮤니티 분야	그래프 데이터베이스 , 그래프 머신러닝 , 네트워크 사이언스
전문분야	그래프 데이터를 활용하고자 하는 초심자를 대상으로 튜토리얼 배포(Graph travel), 그래프 기술에 관심있는 인원들을 대상으로 매주 1회 뉴스레터 발송(Graph omakase), 그래프 학산업 종사자들을 대상으로 인터뷰 진행 및 인사이트 공유(Graph Interview)
글로벌 커뮤니티 활동여부	미디엄 및 트위터를 통해 홍보 및 교류 중 https://medium.com/me/stats/post/6d9177314533

● 커뮤니티 소개

그래프 기술 지식 공유 커뮤니티입니다. 그래프 데이터베이스 , 그래프 머신러닝 , 네트워크 사이언스 3가지 분야가 주 분야입니다. 대중에게 그래프 기술을 전달하기 위해 3가지 콘텐츠 (그래프 오마카세, 그래프 트래블 , 그래프 인터뷰)를 제작하며 교류하고 있습니다.

● 커뮤니티 미션/비전

1. 그래프 기술 대중화
2. 미니 저널, 컨퍼런스 개최를 통한 국내 · 외 인재 간 그래프 기술 교류

Q & A