IT’S WHO YOU KNOW

Graph Mining using Recursive Structural Features

Nicola Di Nardo

Politecnico di Milano

Abstract

Questions:

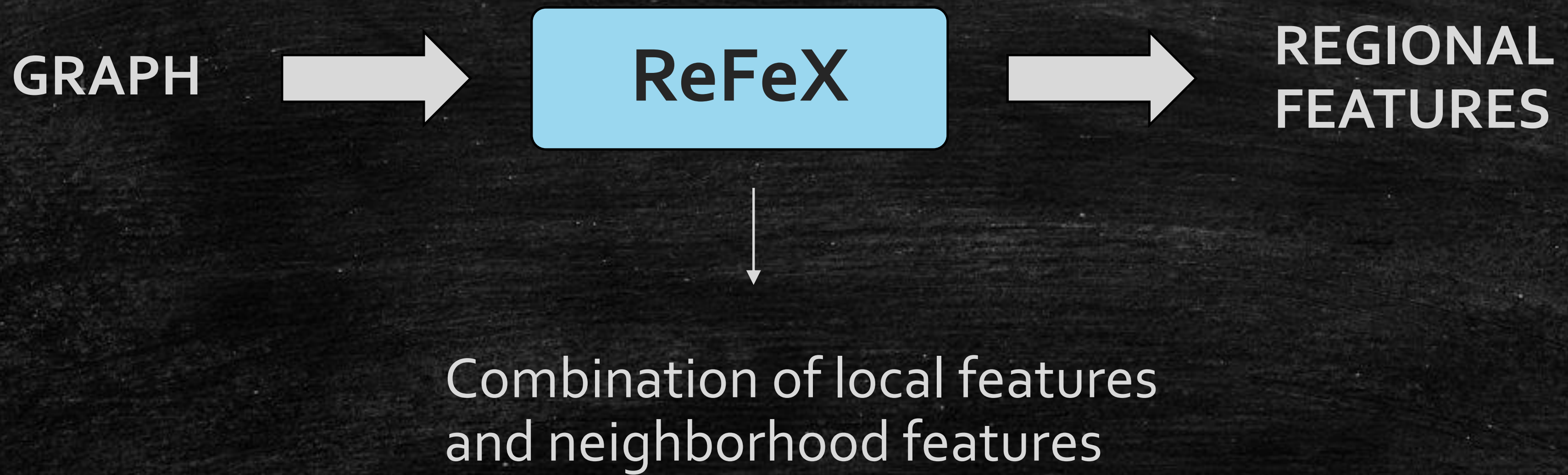
- How can we extract good features from a graph?
- Given two graphs on the same domain, how can we use information in one to make classification in the other?
- If one graph is anonimyzed, how can we use information in one to de-anonimyze the other?

Requirements:

- Effective
- Scalable

ReFeX – Recursive Feature eXtraction

A novel algorithm, that recursively combine local features and neighborhood features and outputs regional features in order to capture behavioural information



Formalization of the problem

Given a graph G as input we have to compute the output as a node-feature matrix F with the following properties:

- Structural
  - The construction of matrix F should not require additional attribute information on nodes and links
  - Topological features only
- Effective
  - It help us to predict node attributes when available
  - It has to be transferable across different graphs

Proposed Algorithm - Definitions

ReFeX aggregates exisisting feature values to generate recursive features

Initial set of features → Structural information

STRUCTURAL ATTRIBUTES

- → Local features
- → Egonet features
- → Recursive features

Local features + Egonet features → Neighborhood features

Neighborhood features + Recursive features → Regional features

- Local features → Essentially node degrees
- Egonet features → Computed on the node ego network
- Recursive features → Any aggregate computed on a feature value among a node's neighbors

Proposed Algorithm - Process

Two steps process:

1. GENERATION
   - Ex. Mean value of the feature degree among all neighbors of a node
   - Not only neighborhood features can be aggregated, also recursive ones

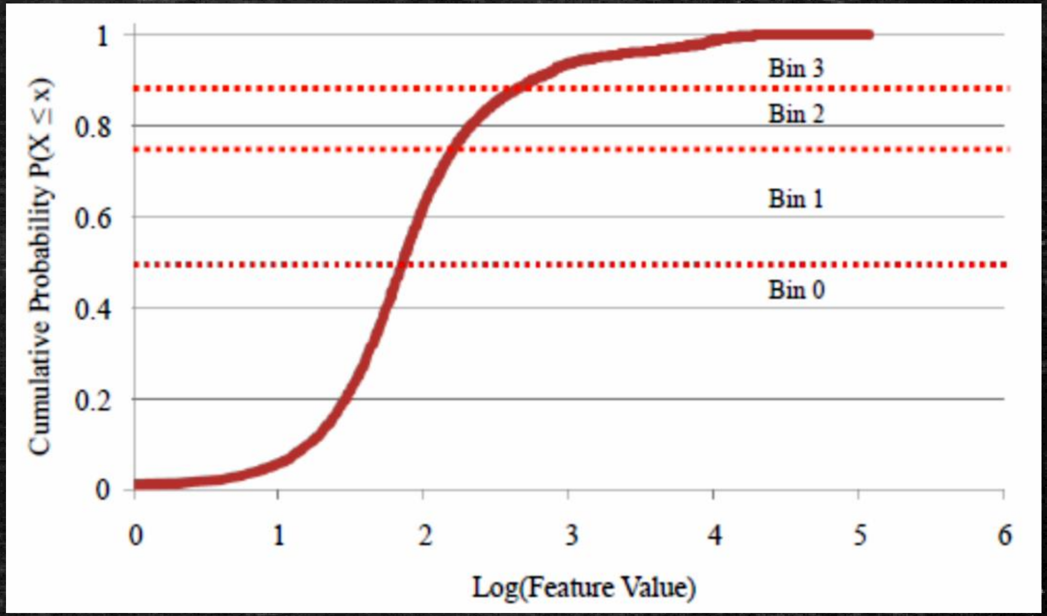
2. PRUNING
   - Impossible to work with an infinite number of features, the generation grows exponentially at each iteration
   - Looks for features that are highly correlated in order to prune at least one of them
   - Different techniques

Proposed Algorithm – Pruning strategy 1

Vertical logarithmic binning

Each feature's values are transformed into vertical logarithmic bins of size p with $0 < p < 1$

1. For feature f_i, the $p*|V|$ lowest f_i values are reassigned value 0
2. Next, p fraction of the remaining nodes are assigned value 1
3. Next, value 2 and so on…



The process is repaeated until every f_i values have been replaced by integers between 0 and $\log_{p-1}|V|$

Proposed Algorithm – Pruning strategy 2

ReFeX looks for pairs of features, after they are generated and binned, that do not disagree more than a threshold s (namely, s-friends features)

How to eliminate redundant features?

1. Build an auxiliary feature graph → Nodes are features; Links are s-friends relations
2. Look for connected component in this graph
3. Replace the entire component by a single feature
   (tipically the «older» one, generated at the earliest iteration)

Proposed Algorithm – Settings & Complexity

SETTINGS

ReFeX requires two parameters:

- p → Fraction of nodes placed in each logarithmic bin → 1 aggressive pruning → 0 runtime complexity
- s → Feature similarity threshold → Tipically uses 0 in the first iteration, then increases over time

COMPLEXITY - two steps

1. Computation of neighborhood features → O(n) for real-world graphs
2. Computation of recursive features (f << n) → Time - O(f*(m + n*f)) → Space - O(m + n*f)

Network Classification - Data

IP-A and IP-B are real network-trace data sets collected roughly one year apart on separate enterprise networks

After some manipulations on the raw data, all the network flows are labeled.
The results are summarized in the table on the right

Note:
There were also other classes in the original data set, but 3 classes (namely, Web, DNS and P2P) made up the dominant traffic type for over the 90% of the labeled hosts

	IP – A1	IP – A2	IP – A3	IP – A4	IP - B
Nodes	81450	57415	154103	206704	181267
labeled	29868	16112	30955	67944	27649
Links	968138	432797	1266341	1756082	1945215
unique	206112	137822	358851	465869	397925
Web	32%	38%	38%	18%	42%
DNS	36%	49%	39%	20%	42%
P2P	32%	12%	23%	62%	16%

- Nodes = IP addresses
- Links = communication between IPs

Network Classification - Classifiers

To test the predictive ability of ReFeX's features, the logForest model described by Gallagher et al. has been used

logForest → Bagged model composed of a set of Logistic Regression classifiers, where each is given a subset of log(f) + 1 of the f total features (500 LR classifiers in this experiment)

wnRN + RL
(as a baseline) → Standard relational neighbor classifier
Memory-based approach with weigthed vote

Within-Network Classification - Methodology

The experiment follows the classical supervised approach in a classification problem

The data set is splitted based on a stratified-class random-sampling:

- **Training Set** → Set of nodes for which we know the labels (from 10% to 90%)
- **Test Set** → Set of unlabeled nodes on which the evaluation is performed

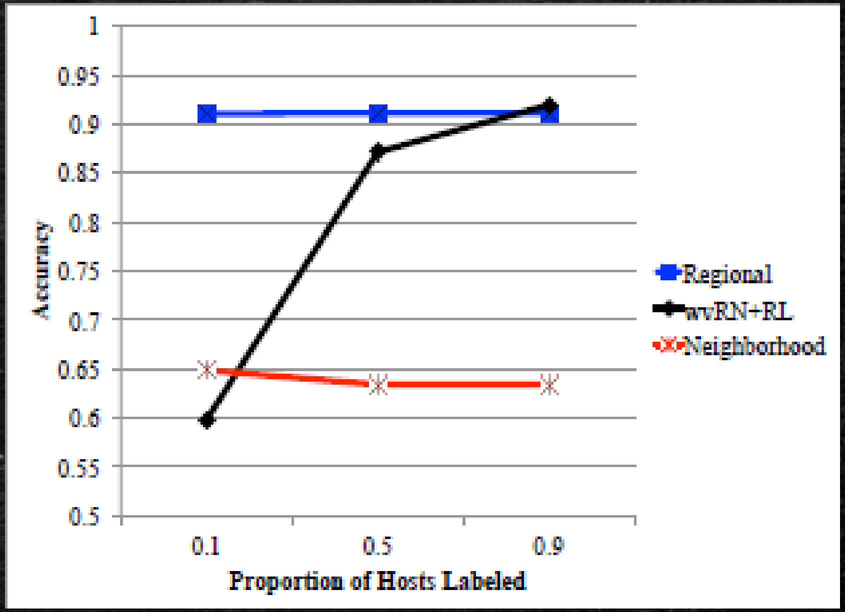
For each proportion of labeled nodes, we run 10 trials and report the average performance

Within-Network Classification - Results

Comparisons:

- wnRN + RL
- logForest on Neighborhood features only
- logForest on Regional features

IP – A3 data set



Performances:

- The Regional classifier outperforms the others almost everywhere
- The Regional and Neighborhood classifiers are less sensitive to the availability of labeled data
- Significant gap in performances when labels are sparse

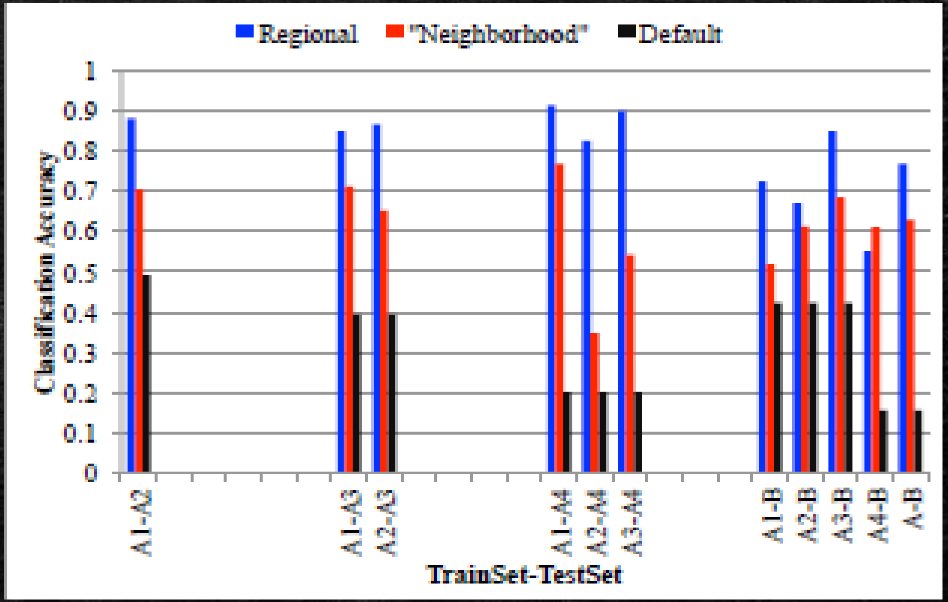
Across-Network Transfer Learning

Methodology:

There are two different graphs of the same domain. The training one completely labeled, while the test one completely unlabeled
The default classifier make prediction based on the most frequent class

Performances:

- The Regional classifier is the best overall performer
- 82% - 91% accuracy on separate days of IP–A
- 77% accuracy training on all days of IP-A and testing on IP-B



Identity Resolution

Now we are facing pairs of networks whose node-sets overlap, computing regional features

GOAL:

Demonstrate that regional features capture meaningful and informative behaviours of nodes

HYPOTHESIS:

Node's feature values will be similar across graphs

POTENTIAL APPLICATION:

Perform «de-anonymization» on social network datasets when external non-anonymized data is available

Identity Resolution - Methodology

We are given two graphs, G_{target} and $G_{reference}$ and a node v_{test} which exists in both graphs

We allow the strategy to guess reference nodes $\langle v_1, \ldots, v_k \rangle$ until it correctly guesses v_{test}

- The score associated with this strategy is k, the number of guesses required to find the node
- The baseline method is to guess at random with an expected score of $|V_{reference}| / 2$

For a given strategy, the guesses are generated in order of increasing Euclidean distance from V_{target} in feature space

To compare the performances we select 1000 nodes in V_{target} with the hisghest degree, and clearly which exists also in $G_{reference}$

Identity Resolution - Data

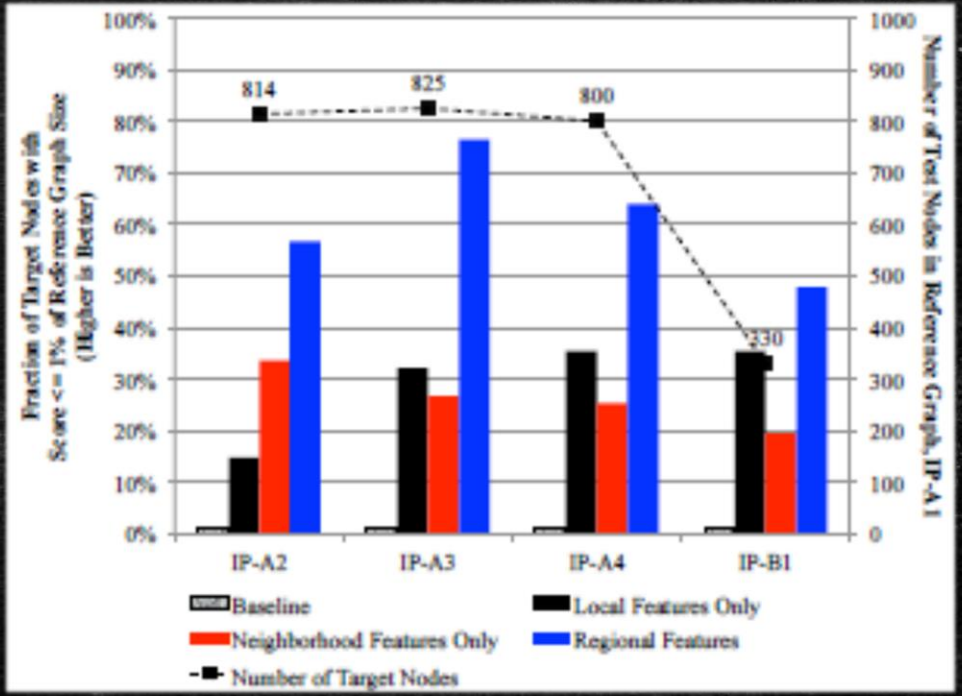
Graph	#Nodes	#Links	Weighted	Directed	#LF	#NF	#RF	#Recursive Iterations
Twitter (T)	465K	845K	Yes	No	3	8	45	6
Twitter (R)	840K	1.4M	Yes	No	3	8	45	-
IP (T)	81K	206K	Yes	Yes	7	22	373	4
IP (R)	57-206K	137-466K	Yes	Yes	7	22	373	-

- Twitter
  - T: who-follows-whom
  - R: who-mentions-whom (first)
- IP
  - T: IP-A1
  - R: IP-A2, IP-A3, IP-A4, IP-B

Identity Resolution – Network Traces

A potential application of this task is to de-anonymize a network trace where IP addresses are hidden observing a non-anonymized enterprise trace

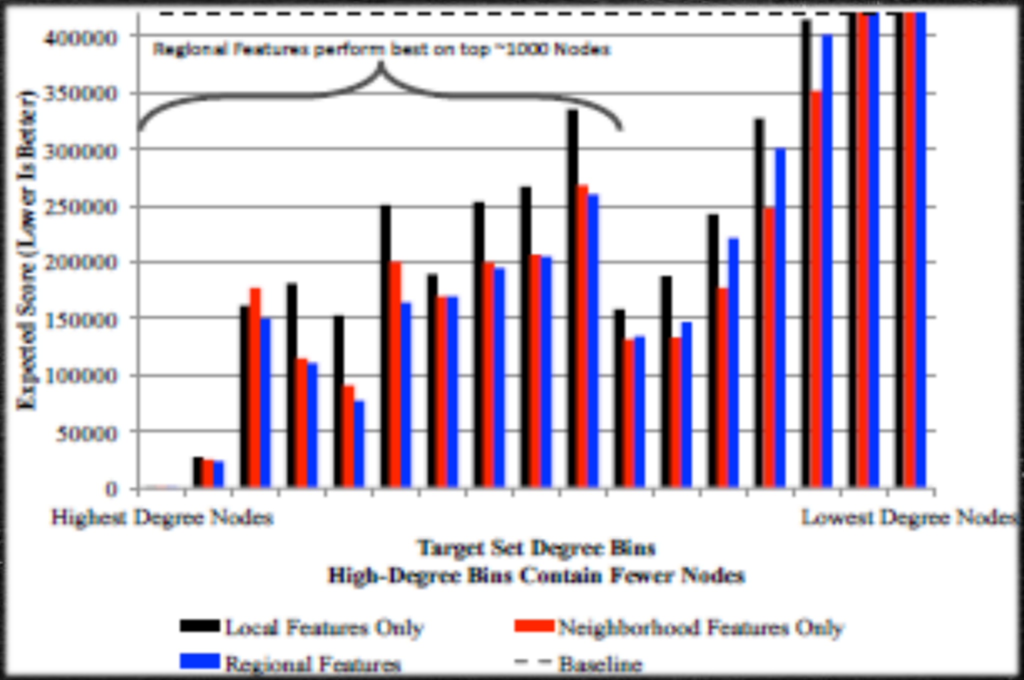
- Regional features dominate the performance of other strategies in all tests
- Over 45% of S_{test} scoring in the top 1% of the size of $G_{reference}$
- There is a small subset of test nodes for which Regional perform very poorly (near baseline)



Identity Resolution - Twitter

Success at this task indicates that one could de-anonymize a social network by using public available text data, so long as usernames can be parsed from text

- Regional features outperforms other strategies in the first ten bins (highest degree nodes)
- For lowest degree nodes, all strategies perform worse than the baselin
    - Fewer observed behavior to leverage
    - Many more similar nodes



Conclusions

- We described a novel algorithm ReFeX, which extracts regional features from nodes based on their neighborhood connectivity
- These regional features capture information in terms of the kind of nodes to which a given node is connected as opposed to the identity of those nodes
- We showed that ReFeX is effective and scalable in various graph mining tasks including within- and across-network classification and identity resolution tasks

Thank you for your attention!

References 1

- L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graps
- R. Albert, H. Jeong and A.-L. Barabasi. Diameter of the world wide web. Nature
- A.-L. Barabasi, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. Physics A: Statistical Mechanics and its Applications
- J. Blitzer, M. Dredze, and F. Pereira. Biograohies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification
- D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. SIAM Int. Conf. on Data Mining, Apr. 2004
- W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In ICML
- M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationship of the internet topology. SIGCOMM
- H. Fei and J. Huan. Structure feature selection for graph classification. In CIKM, 2008
- G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. IEEE Computer, Mar. 2002
- B. Gallagher and T. Eliassi-Rad. Leveraging label-independentfeatures for classification in sparsely labeled networks: An empirical study. Lecture Notes in Computer Science, 2010

References 2

- B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos. Using ghost edges for classification in sparsely labeled networks
- J. Gao, W.Fan, J. Jiang and J. Han. Knowledge transfer via multiple model local structure mapping. In KDD, 2008
- N. Ghamrawi and A. McCallum. Collective multi-label classification. In CIKM, 2005
- M. Girvan and M. E. J. Newman. Community structure in social and biological networks. PNAS, 2002
- J. He, Y. Liu, and R. D. Lawrence. Graph-based transfer learning. In CIKM, 2009
- K. Henderson, T. Eliassi-Rad, C. Faloutsos, l. Akoglu, L. Li, K. Maruhashi, B. A. Prakash, and H. Tong. Metric forensics: a multi-level approach for mining valatile graphs. In KDD,2010
- J. M. Kleinberg, R. Kumar, P. Raghavan. S. Rajagopalan, and A. S. Tomkins. The Web as a graph:Measurements, models and methods. Lecture Notes in Computer Science, 1999
- X. Kong and P.S. Yu. Multi-label feature selection for graph classification. Data Mining, IEEE International Conference on, 2010
- S. – I. Lee, V.Chatalbashev, D. Vickrey, and D. Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In ICML, 2007

Reference 3

- J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph over time: densification laws, shrinking diametes and possible explanations. In Proc of ACM SIGKDD, 2005
- R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In KDD, 2010
- C. Liu, X. Yan, H. Yu, J. Han, and P. S. Yu. Mining behaviour graphs for backtrace of non-crashing bugs. In SDM, 2005
- S. A. Macskassy and F. Provost. A simple relational classifier. In Proceedings of the Second Workshop on Multi-Relational Data Mining at KDD, 2003
- M. E. J. Newman. Power laws, pareto distributions and zipf's laws. Contemporary Physics, 2005
- C. C. Noble and D. J. Cook. Graph-based anomaly detection. In KDD, 2003
- H.Tong, B.A. Prakash, C. E. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau. On the vulnerability of large graphs. In ICDM, 2010
- I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann Publishers Inc., 2005
- L. Yao, S. Riedel, and A. McCallum. Collective cross-document relation extraction without labelled data. In EMNLP, 2010
- J. Zhang, Z. Ghahramani, and Y. Yang. Learning multiple related tasks using talent independent component analysis. In NIPS, 2005