RSAConference2018
San Francisco | April 16 – 20 | Moscone Center

MATTERS NOW

SESSION ID: MLN-F01

FIGHTING MALWARE WITH GRAPH ANALYTICS: AN END-TO-END CASE STUDY

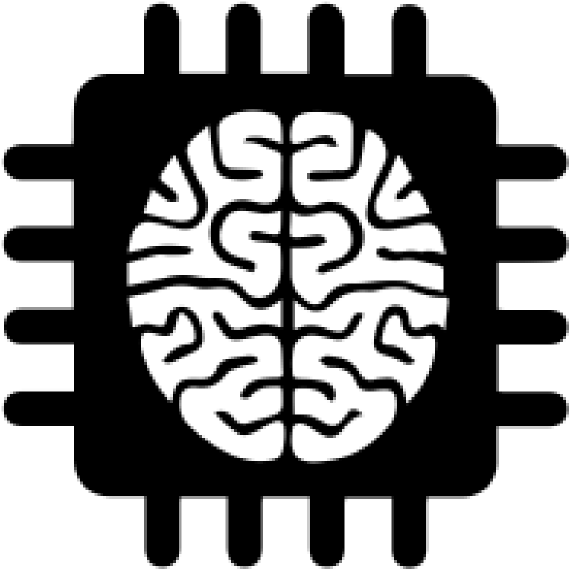
Mayana Pereira

Data Scientist
Infoblox Inc.

MACHINE LEARNING

Profile picture →

Posts →



→ Age

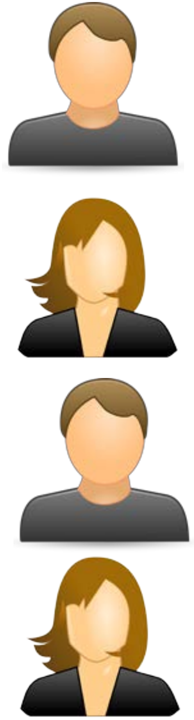
→ Gender

→ Personality

→ Political Preferences

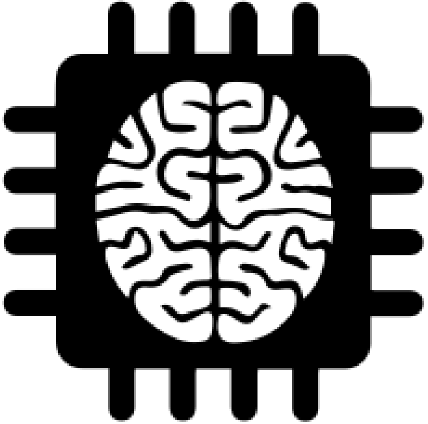
→ Advertising

MACHINE LEARNING



~~Profile picture~~ →

~~Posts~~ →

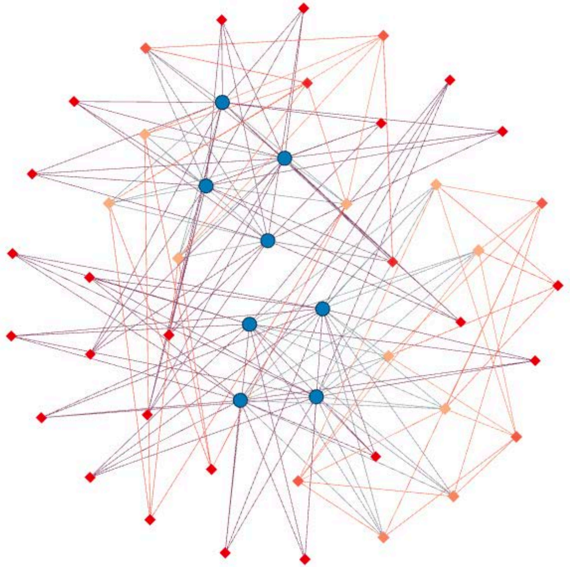


→ Fake Users

→ Influencers

MACHINE LEARNING

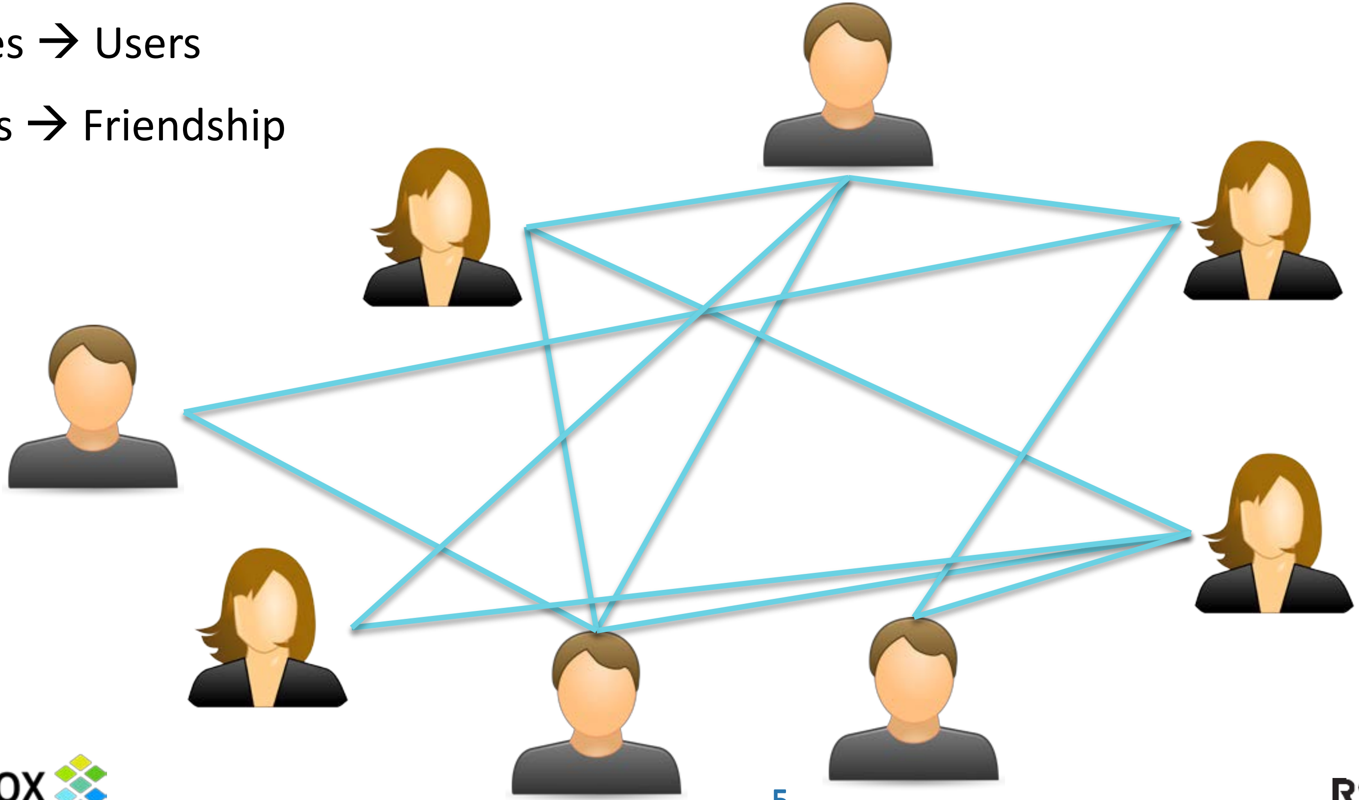
Relations between users → **Graphs**



GRAPH DEFINITION

Nodes → Users

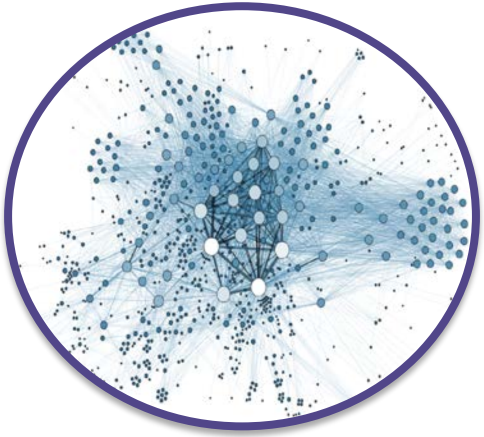
Edges → Friendship



GRAPH-BASED ANALYTICS

- A powerful tool for modeling, structuring and understanding relationships among people, devices, information entities.
- Graph mining provides an insightful representation
  - Interdependent instances
  - Long-range relations
  - Node/Edge attributes (data complexity)
  - Hard to fake/alter (adversarial robustness)
- Security-related applications:
  - Exposing Terrorist cells
  - Botnet detection
  - Reputation propagation of IPs

OUR GOAL IS…

Learn	+	Apply	=	Obtain
				
How graphs and graph mining can be used to solve security problems. Intuitive and easy to understand approaches.		Identify the data related problems and how to describe them through graphs.		Graph-based machine learning models that complements and can outperform traditional techniques.

SUMMARY

#RSAC

1. Study Case: Malware Detection using Graph Mining
   a) Dictionary-DGA Problem
   b) Graph-based analytics to extract malware dictionaries from DNS traffic

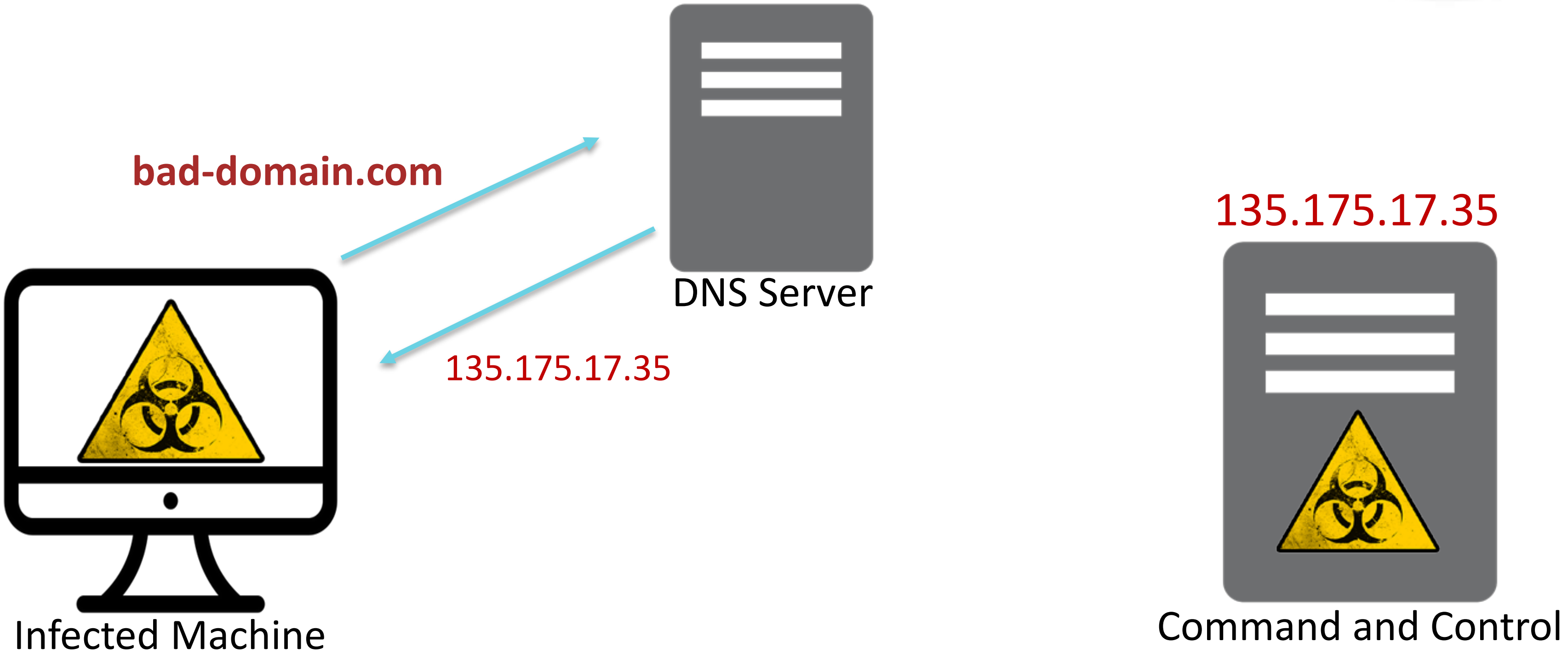
2. Graph Mining techniques that are easy to visualize and interpret
   a) EgAnomaly Detection
   b) Open Source tools for data exploration & visualization tools

SUMMARY

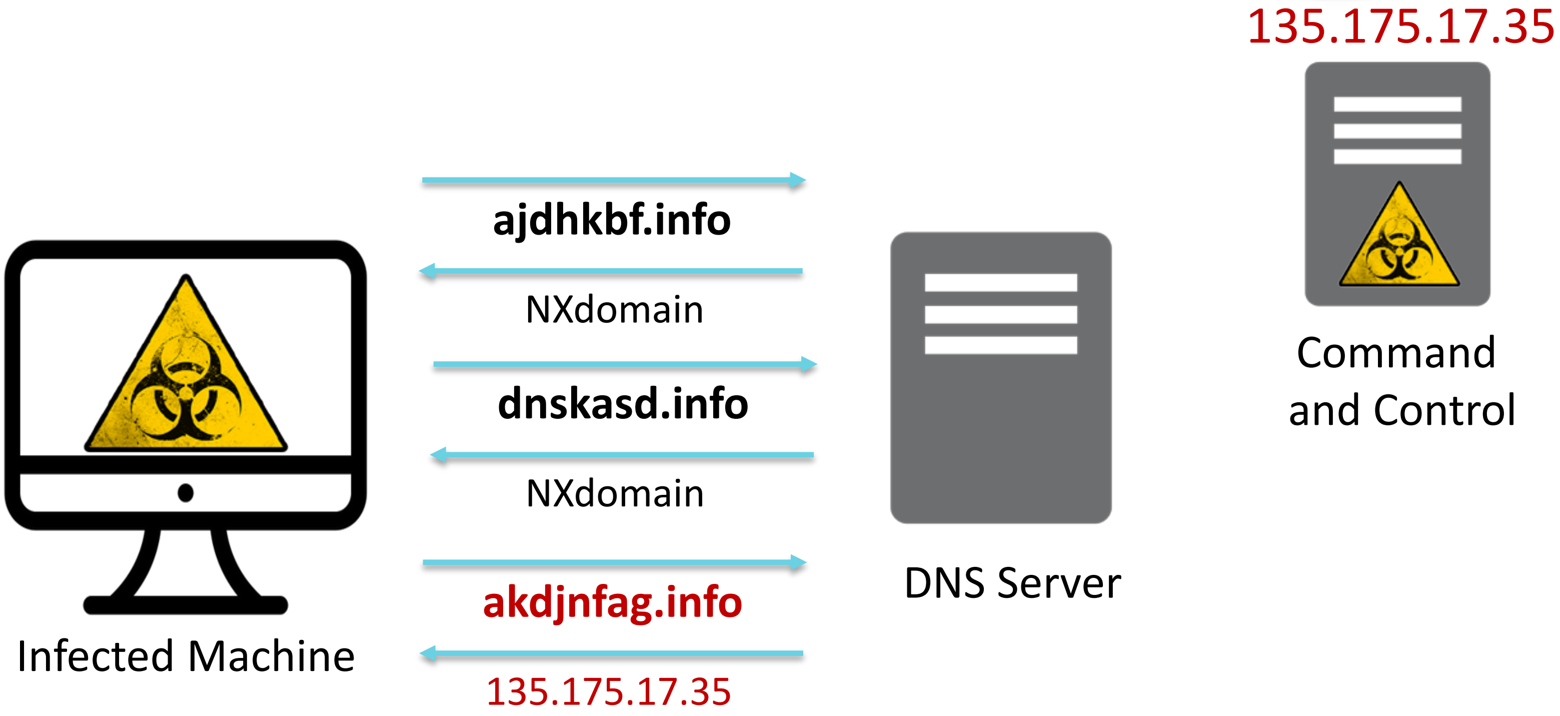
1. Study Case: Malware Detection using Graph Mining
   a) Dictionary-DGA Problem
   b) Graph-based analytics to extract malware dictionaries from DNS traffic

2. Graph Mining techniques that are easy to visualize and interpret
   a) Anomaly Detection
   b) Open Source tools for data exploration & visualization tools

MALWARE COMMUNICATION

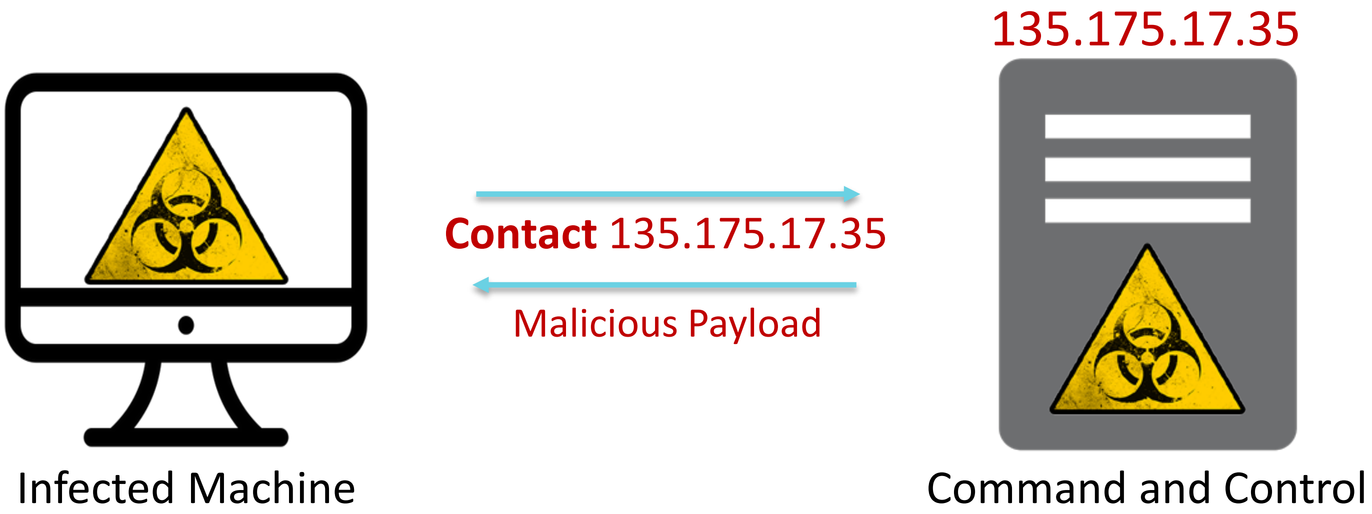


DOMAIN GENERATION ALGORITHMS

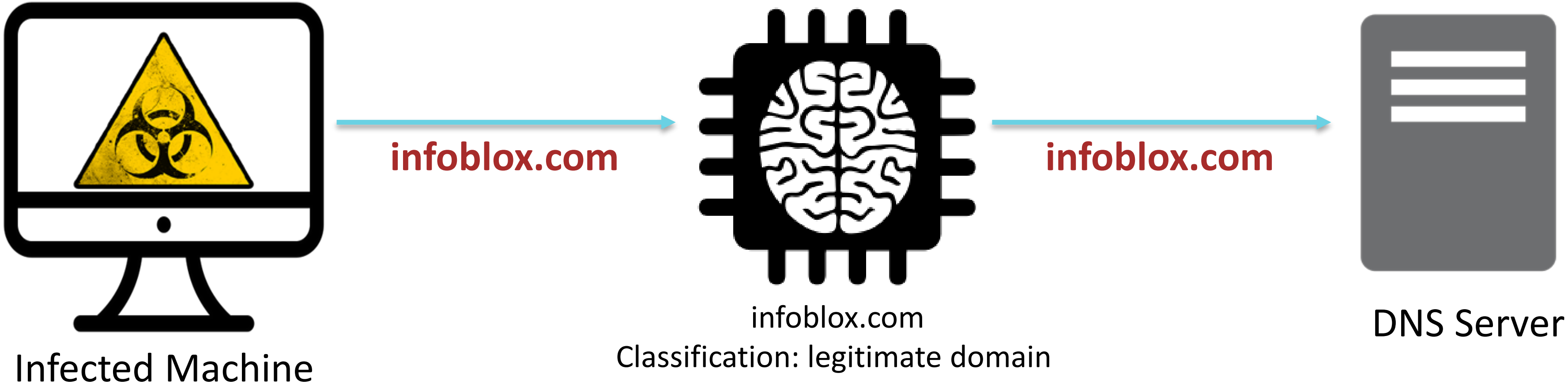


DOMAIN GENERATION ALGORITHMS

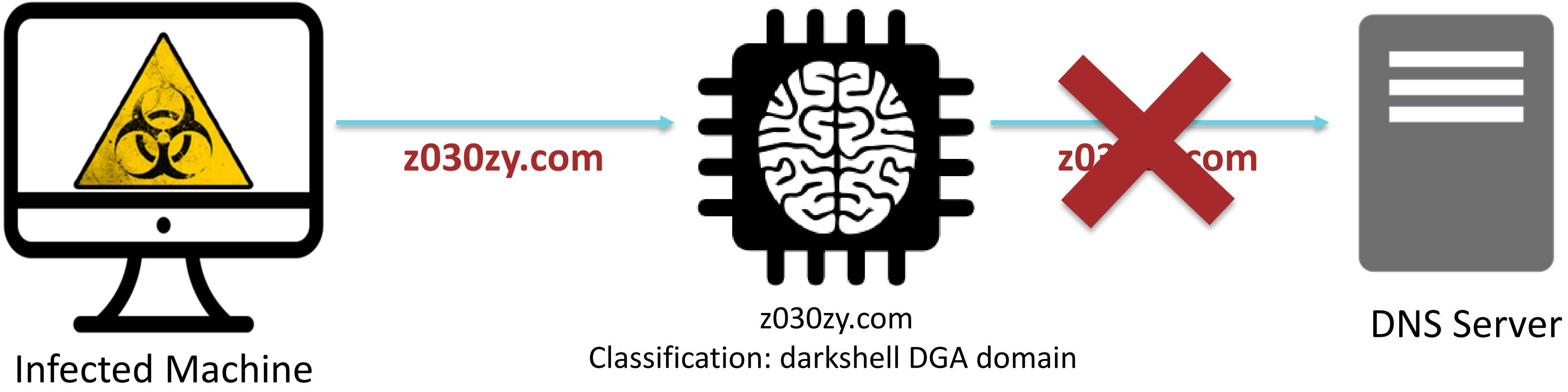
#RSAC



DOMAIN GENERATION ALGORITHMS



DOMAIN GENERATION ALGORITHMS



katherinelangford.net

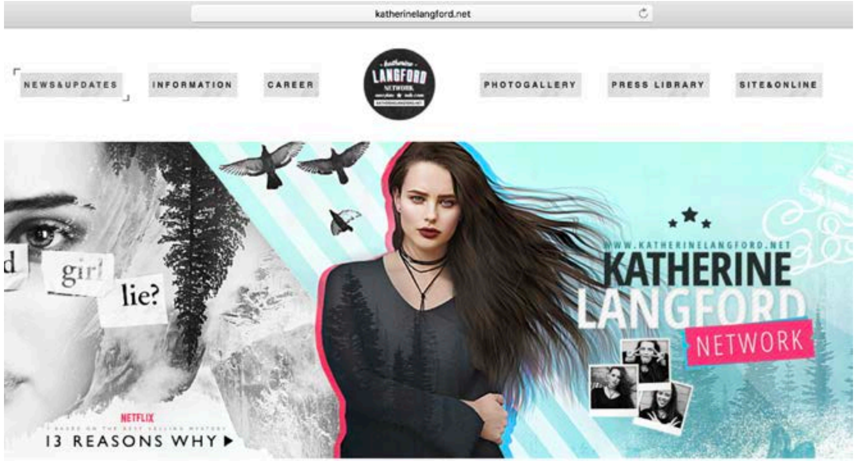
X

kyt6ea4ak4bvo35lrw.net

Can you tell which one is a DGA domain and which one is a Hollywood actress website?

DGA DETECTION

katherinelangford.net



kyt6ea4ak4bvo35lrw.net

Rovnix Malware DGA domain

HOW DIFFICULT IS THE PROBLEM?

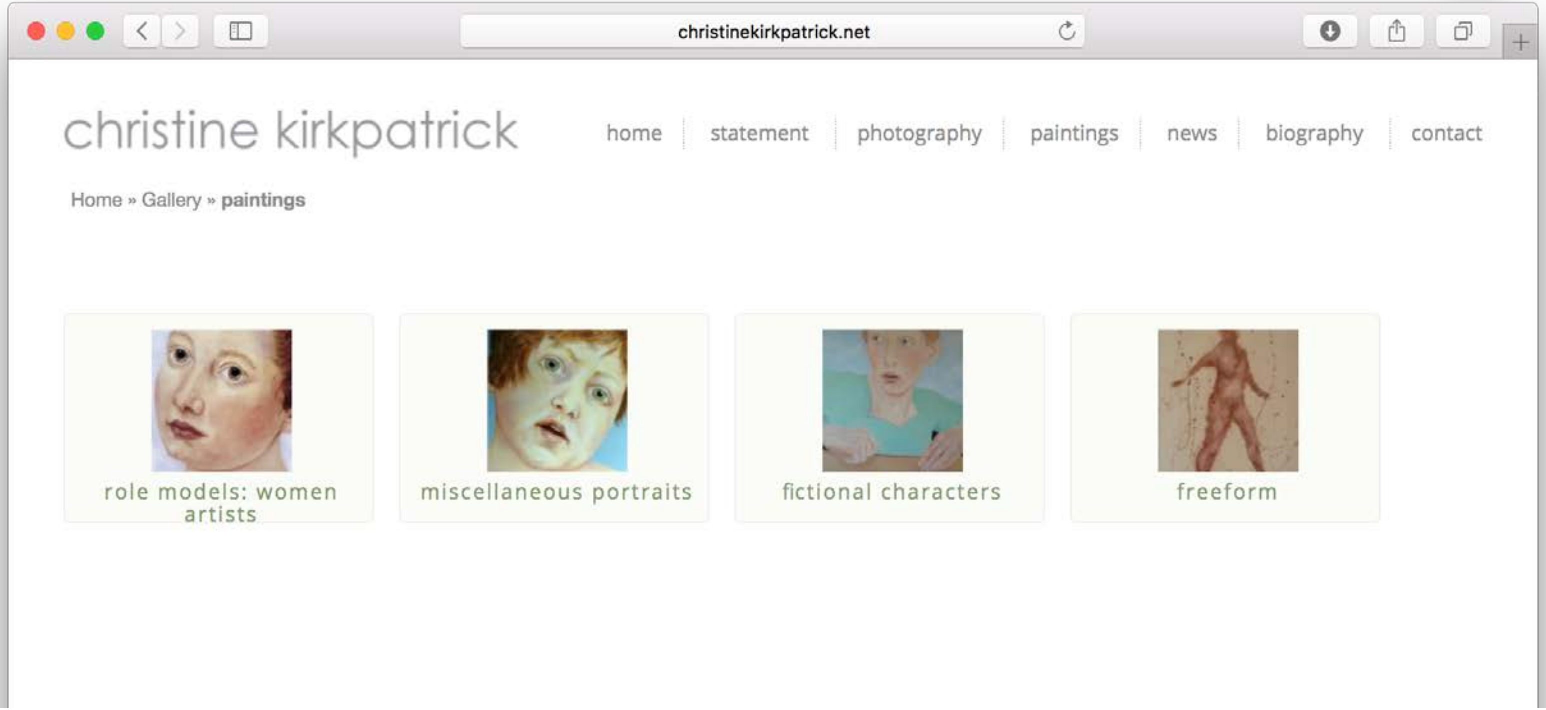
christinepatterson.net

X

christinekirkpatrick.net

Can you tell which one is a DGA domain and which one is an online art gallery?

IT IS A DIFFICULT PROBLEM



IT IS A DIFFICULT PROBLEM

christinepatterson.net

Is a DGA domain!

Suppobox Malware Family

OBSERVE GROUP OF DOMAINS

hurt

wishwear.net
joinhurt.net
wishhurt.net
deadtold.net
rocktold.net
deadfind.net
rockfind.net
deadwear.net
rockwear.net
deadhurt.net
rockhurt.net
wrongtold.net
madetold.net
wrongfind.net
madefind.net
wrongwear.net
madewear.net
wronghurt.net
madehurt.net

stephaniesackville.net
charlottehoneycutt.net
stephaniehoneycutt.net
charlottefairchild.net
stephaniefairchild.net
kimberlynpettigrew.net
glanvillepettigrew.net
kimberlynsackville.net
glanvillesackville.net
kimberlynhoneycutt.net
glanvillehoneycutt.net
kimberlynfairchild.net
glanvillefairchild.net
jessaminepettigrew.net
genevievepettigrew.net
jessaminesackville.net
genevievesackville.net
jessaminehoneycutt.net

THE WORDS REPEAT

wishwear.net
joinhurt.net
wishhurt.net
deadtold.net
rocktold.net
deadfind.net
rockfind.net
deadwear.net
rockwear.net
deadhurt.net
rockhurt.net
wrongtold.net
madetold.net
wrongfind.net
madefind.net
wrongwear.net
madewear.net
wronghurt.net
madehurt.net

stephaniesackville.net
charlottehoneycutt.net
stephaniehoneycutt.net
charlottefairchild.net
stephaniefairchild.net
kimberlynpettigrew.net
glanvillepettigrew.net
kimberlynsackville.net
glanvillesackville.net
kimberlynhoneycutt.net
glanvillehoneycutt.net
kimberlynfairchild.net
glanvillefairchild.net
jessaminepettigrew.net
genevievepettigrew.net
jessaminesackville.net
genevievesackville.net
jessaminehoneycutt.net

honeycutt

SUMMARY

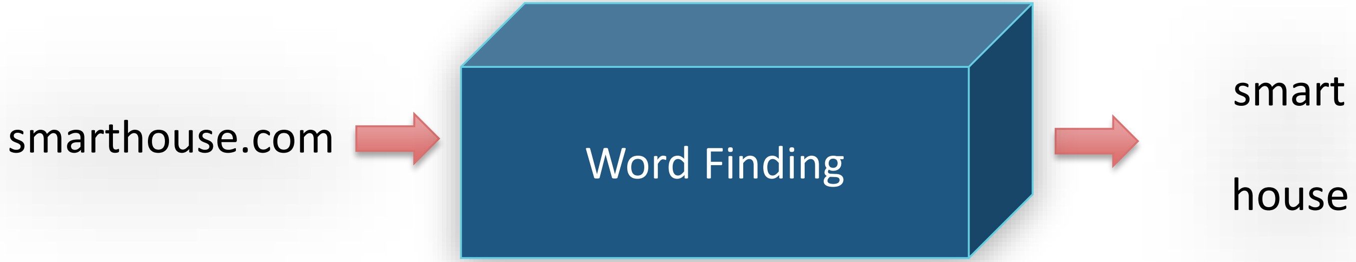
1. Study Case: Malware Detection using Graph Mining
   a) Dictionary-DGA Problem
   b) Graph-based analytics to extract malware dictionaries from DNS traffic

2. Graph Mining techniques that are easy to visualize and interpret
   a) Egonet analysis for Anomaly Detection
   b) Open Source tools for data exploration & visualization tools

BUILDING THE GRAPH

doghouse.com → Word Finding → dog, house

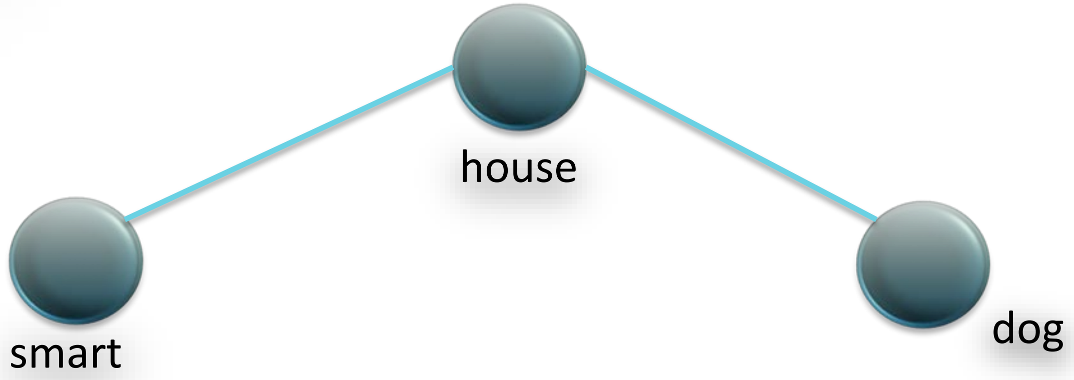
BUILDING THE GRAPH



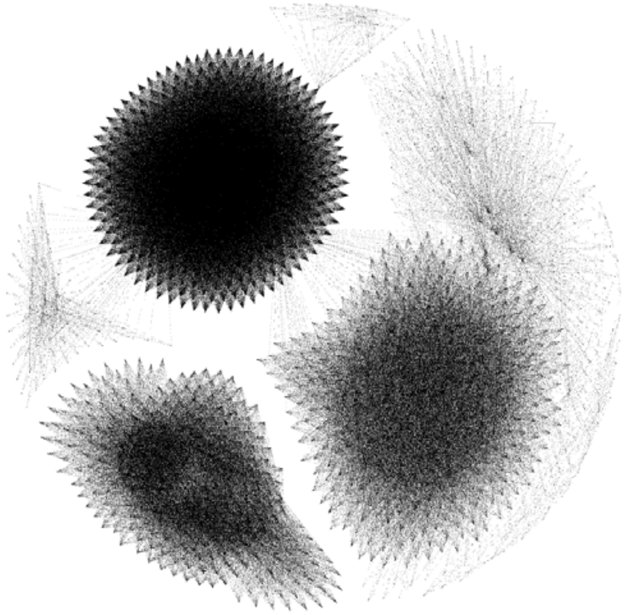
BUILDING THE GRAPH

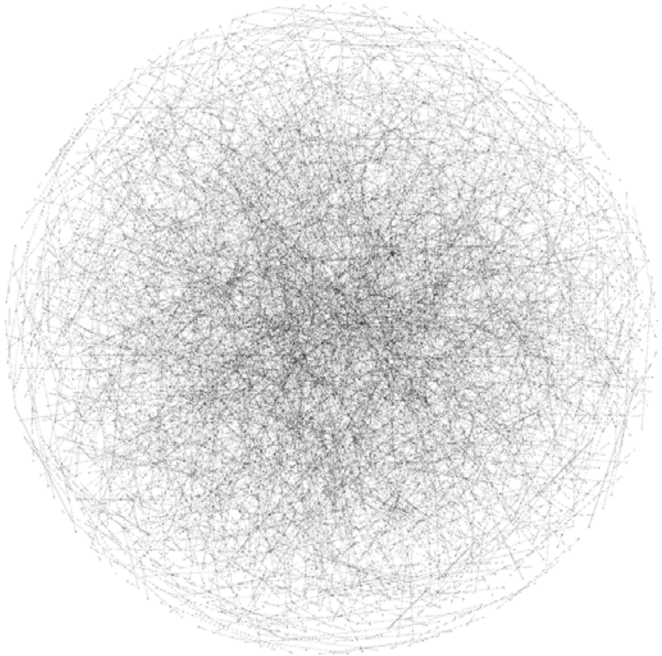
doghouse.com

smarthouse.com



DGA WORDS CONNECT DIFFERENTLY





DETECTING DICTONARIES

1

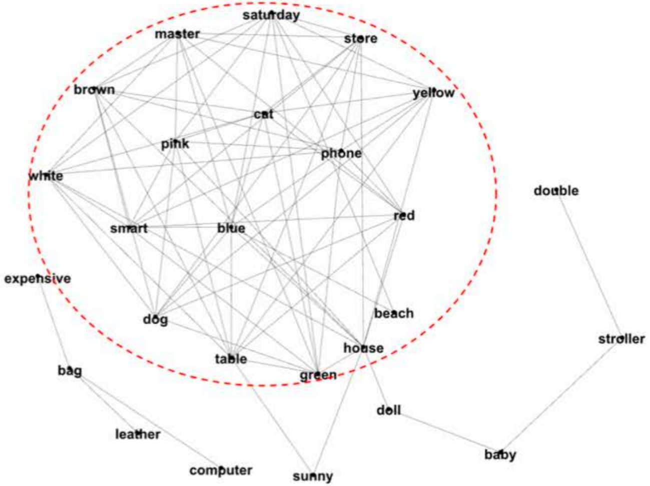
housedoll.com
babydoll.com
babystroller.com
doublestroller.com
housesunny.com
tablesunny.com
saturdaybeach.com
computerbag.com
expensivebag.com
leatherbag.com
...

housewhite.com
houseblue.com
housered.com
dogred.com
doggreen.com
dogbrown.com
tablewhite.com
tablestore.com
masterred.com
phonewhite.com
...

Dictionary AGDs

Legitimate Domains

2



3

Malware Dictionary

'house', 'dog', 'smart', 'table', 'cat', 'master', 'phone', 'red', 'white', 'blue', 'green', 'brown', 'pink', 'yellow', 'store', 'saturday'

WE EXTRACT THE DICTIONARIES WITHOUT REVERSE ENGINEERING EFFORTS!

MALICIOUS REGION IDENTIFICATION

ID	D_{mean}	D_{max}	C	C_V	ASPL	Label
ID_1	7.16	16.0	63	2.62	1.84	True
ID_2	6.91	16.0	60	2.50	1.86	True
…	…	…	…	…	…	
ID_N	3.54	80.7	20	1.7	3.78	False

Features

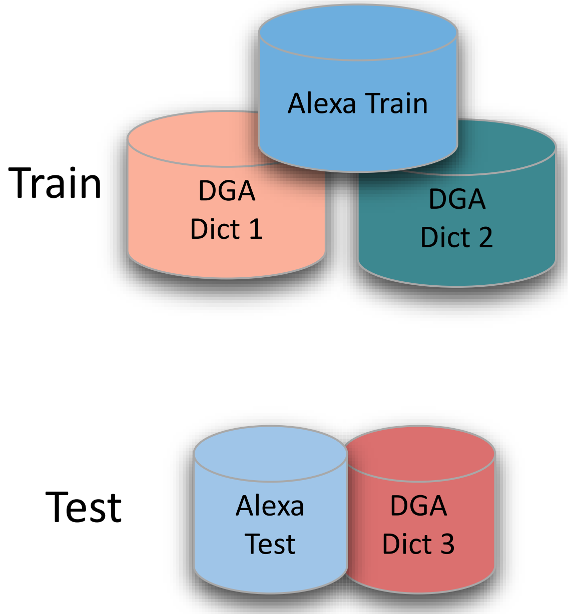
D_{mean}: Average node degree
D_{max}: Maximum node Degree
C: Cardinality of basis of cycles of G
C_V: Average cycles per node
ASPL: Average Shortest Path Length

GRAPH MINING IS POWERFUL

#RSAC

Unbalanced Dataset: DGA domains are less than 1%

120K Benign Domains
1020 DDGA domains



	# of words used by DGA	# of detected words	Recall	FPR
Round 1	92	92	1	0
Round 2	70	64	0.91	0
Round 3	80	80	1	0

HIGH DETECTION RATE

#RSAC

Classification Results

	Round 1			Round 2			Round 3		
Model	Precision	Recall	FPR	Precision	Recall	FPR	Precision	Recall	FPR
WordGraph	1	1	0	1	0.96	0	1	1	0
Random Forest (Baseline)	0.056	0.009	10^{-3}	0.031	0.006	10^{-3}	0.0	0.0	10^{-3}

Infoblox

RSAConference2018

SUMMARY

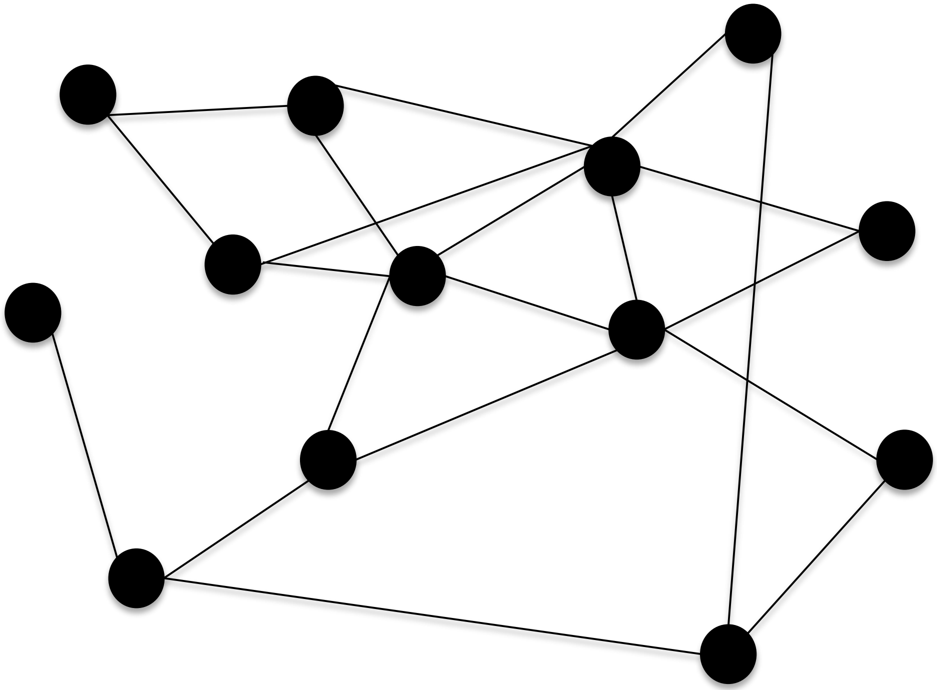
1. Study Case: Malware Detection using Graph Mining
   a) Dictionary-DGA Problem
   b) Graph-based analytics to extract malware dictionaries from DNS traffic

2. Graph Mining techniques that are easy to visualize and interpret
   a) Anomaly Detection
   b) Open Source tools for data exploration & visualization tools

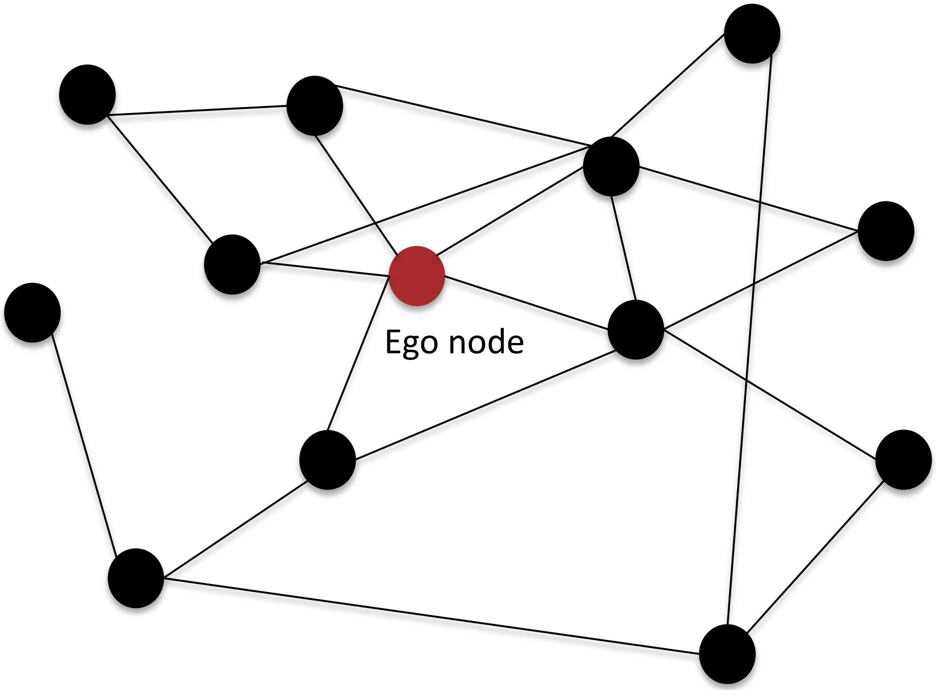
ANOMALY DETECTION

- How can we find anomalies in a graph?
- Anomalous behavior could signify
  - Fraud
  - Network intrusion
  - Electronic auction fraud
  - Etc.

EGONET

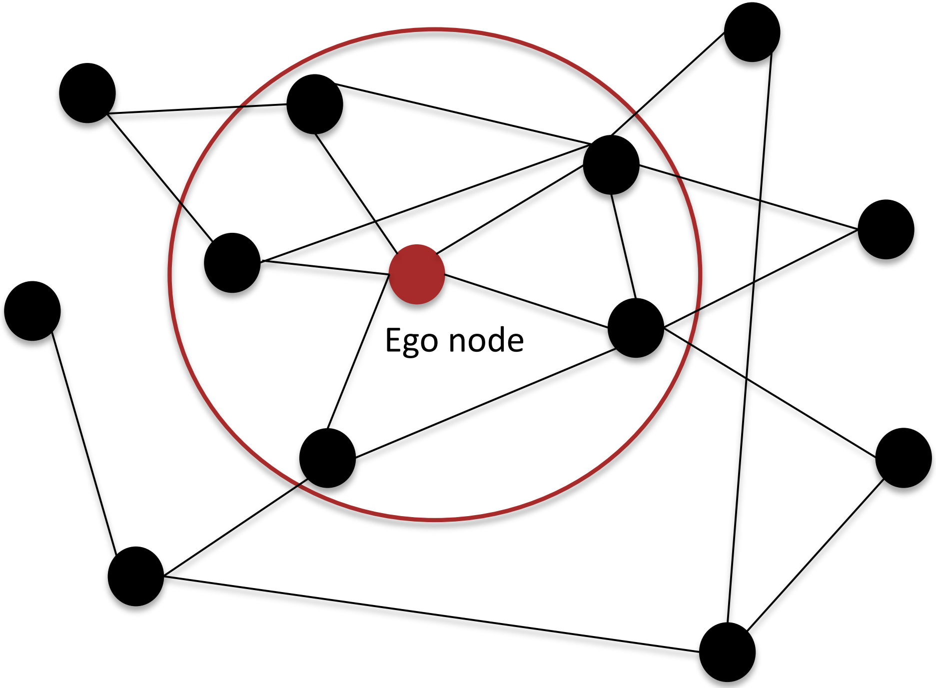


EGONET



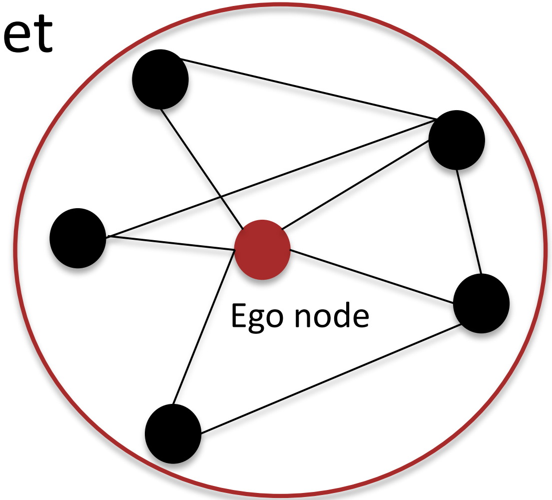
EGONET

Neighborhood around a node



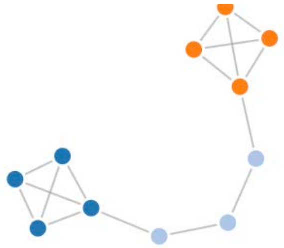
EGONET

Egonet



python

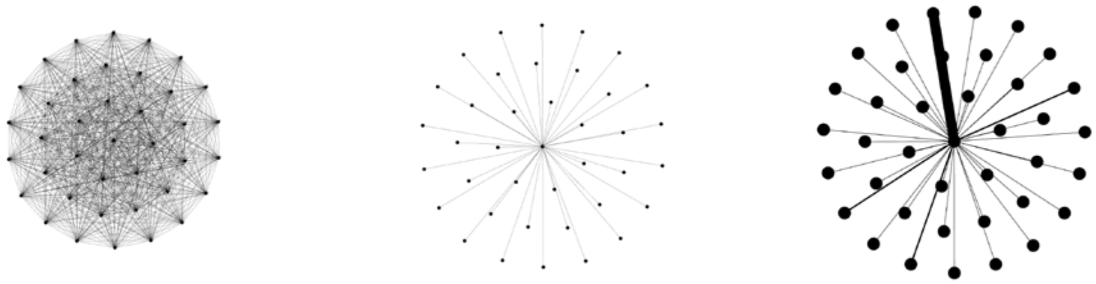
NetworkX



`ego_graph(G, n, radius=1, center=True, undirected=False, distance=None)`

ANOMALY DETECTION

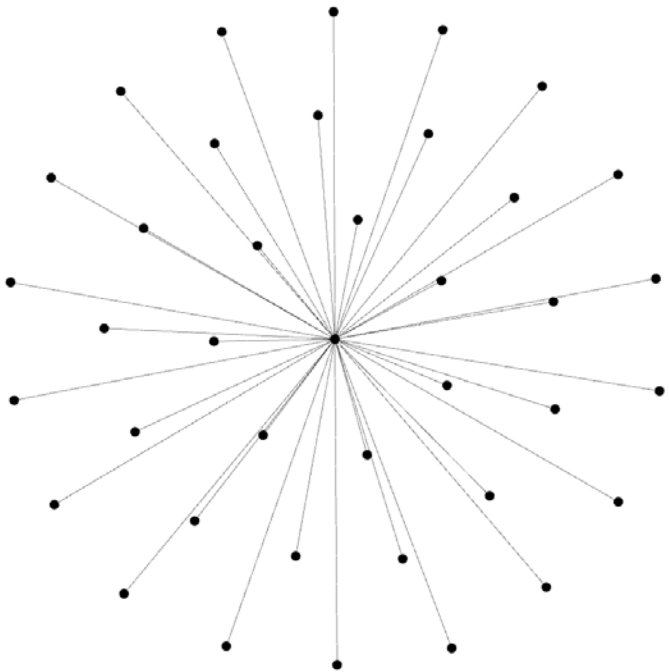
- Finding Anomalous nodes in graphs by analyzing their Egonets
- Egonet behaviors that can indicate anomaly



ANOMALY DETECTION

Spam email accounts

Spamming email accounts usually send emails in a robotic fashion to many accounts, that are unrelated to each other.



ANOMALY DETECTION

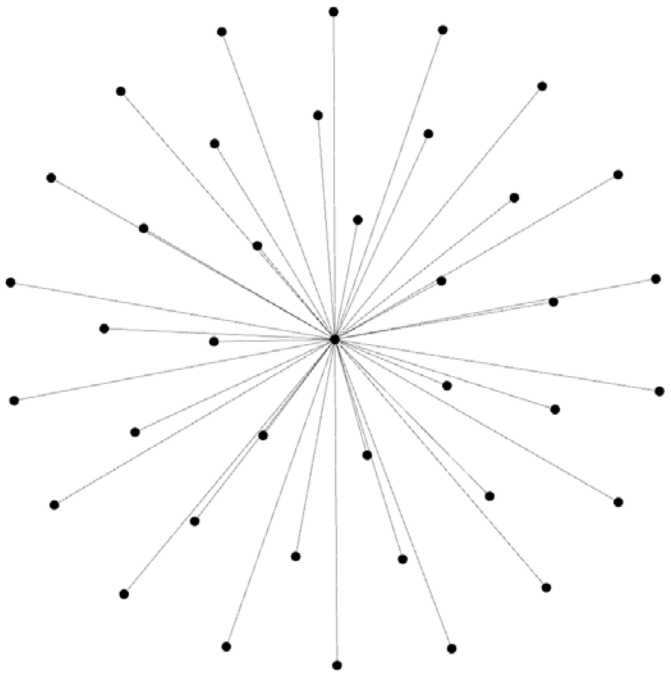
Finding a Star Egonet (Near-Star)[2]

- $E = N^{\alpha}$, $\alpha < 1.1$
  - E: number of edges in a Egonet
  - N: Degree of an egonet (numer of nodes)

NetworkX

```
Graph.degree
```

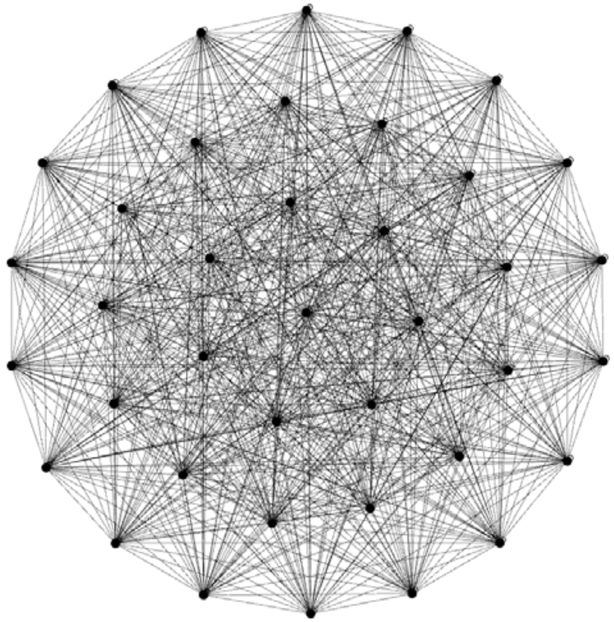
```
Graph.number_of_edges
```



ANOMALY DETECTION

IRS Fraud – Phony tax returns

In a graph model in which each tax form is a node, and the existence of an edge is based on tax form similarity, i.e. any two tax forms that are suspiciously similar have a relationship (edge), any sufficiently large clique identifies a cluster worth studying.

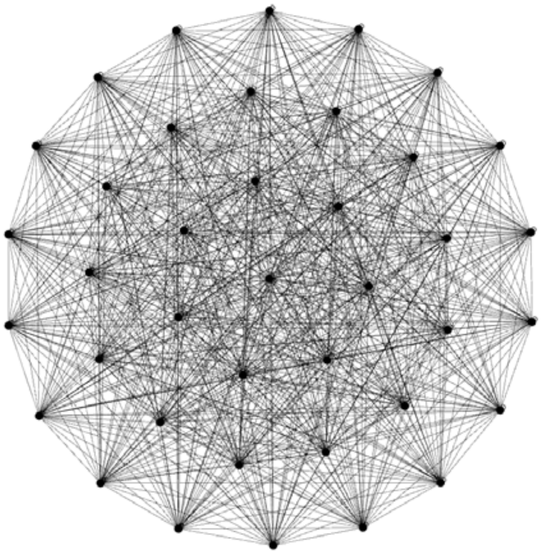


Finding a group of nodes with clique or near clique egonets → Fraud.

ANOMALY DETECTION

Finding a Clique Egonet (Near-Clique)

- $E = N^{\alpha}$, $\alpha > 1.7$
  - E: number of edges in a Egonet
  - N: Degree of an egonet (number of nodes)



NetworkX

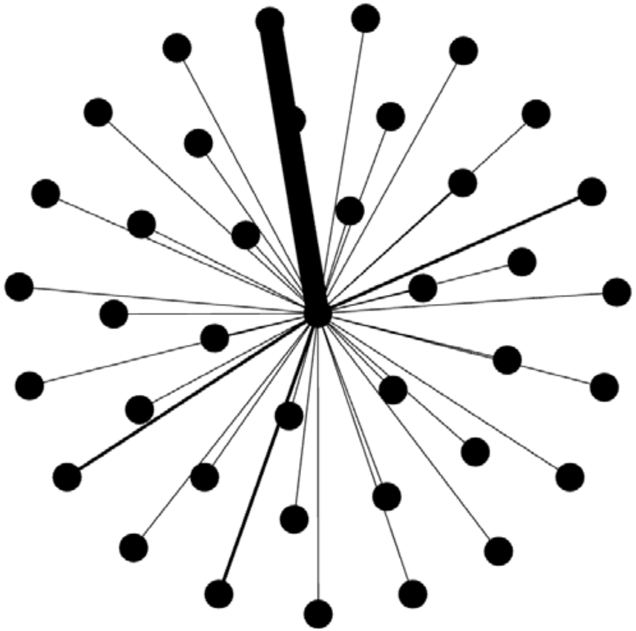
```
Graph.degree
```

```
Graph.number_of_edges
```

ANOMALY DETECTION

Fraud in Credit Card Transactions

An unusual high transaction can indicate stolen or cloned credit card. This is represented by a dominant edge in an Egonet.

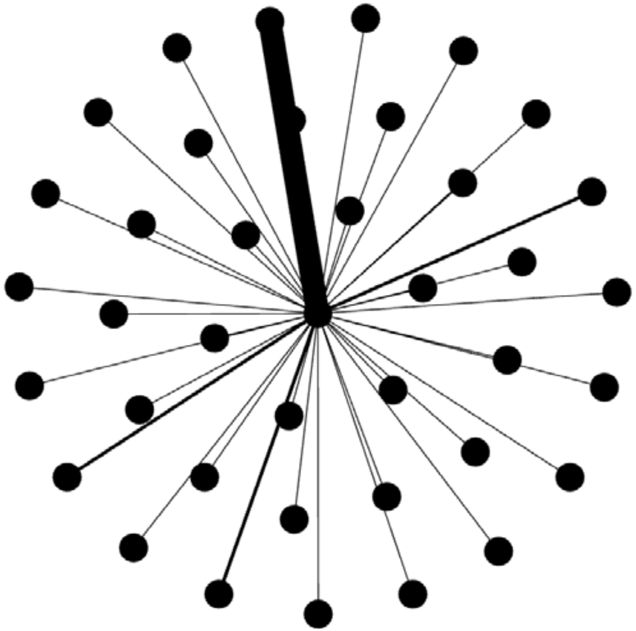


ANOMALY DETECTION

Finding a Heavy Edge

- $\boldsymbol{\lambda} = W^{\alpha}$, $\alpha > 0.98$
  - $\boldsymbol{\lambda}$: Principal Eigenvalue of an Egonet
  - W: Egonet weight (sum of the weights of all edges)

`MultiDiGraph.get_edge_data(u, v, key=None, default=None)`



SUMMARY

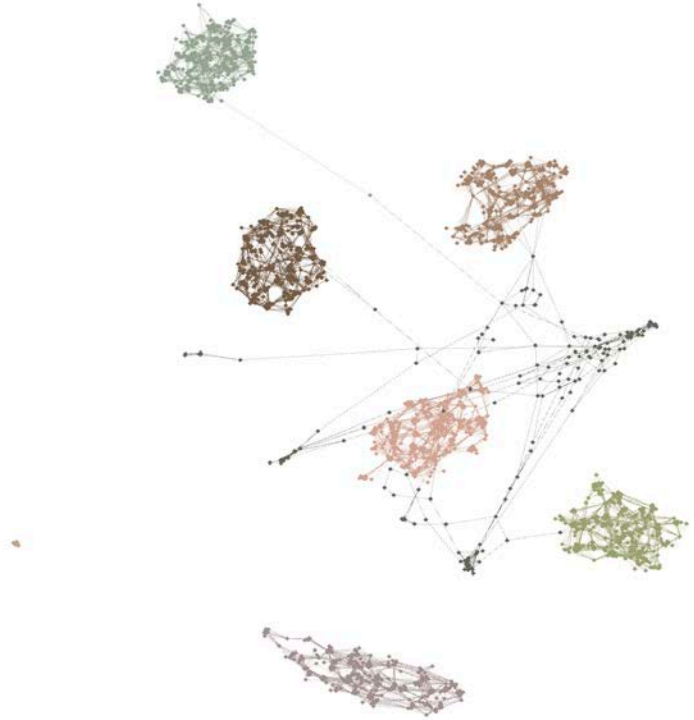
1. Study Case: Malware Detection using Graph Mining
   a) Dictionary-DGA Problem
   b) Graph-based analytics to extract malware dictionaries from DNS traffic

2. Graph Mining techniques that are easy to visualize and interpret
   a) Anomaly Detection
   b) Open Source tools for data exploration & visualization tools

COMMUNITY DETECTION

Community structure means that the network may be clustered to sets of nodes, each of which relatively densely-interconnected internally, with relatively few connections between sets.

Used in study of many fields such as social networks, social studies, computer networks, etc.



COMMUNITY DETECTION

- Useful Tool for Graph Exploration
  - Can we find communities in a certain graph structure?
- Detecting Malicious Group Behavior [5]
  - Problem of Detecting spammers in Email Service Providers (ESPs)
    - Performs better than Content-based filters.
    - Early detection of spamming accounts.

python-louvain 0.10

Louvain algorithm for community detection

Gephi
makes graphs handy

TOOLS FOR GRAPH ANALYSIS

- How to Explore?
  - Python's Libraries NetworkX and python-Louvain are easy to start.
    - https://networkx.github.io
    - https://github.com/taynaud/python-louvain

- How to visualize?
  - Gephi is a great Graph visualization tool that allows to apply community detection while visualizing the data – this is great for data exploration!
    - https://gephi.org

TODAY WE LEARNED…

- Representing information in the right way is the key to find what you want.
- Graphs are a powerful representation for highlighting group structures.
- They are a powerful ally in: fraud detection and intrusion detection and group classification problems.
- There are powerful open source tools for graph mining and analysis.

APPLY WHAT YOU LEARNED TODAY!

- **Next week you should:**
  - Identify problems that can be represented and explored by graph analytics.
  - Think of unusual node relations when building graphs, just like our study case example.
- **In the first three months following this presentation you should:**
  - Identify which approach is best suitable for the problem you are trying to solve. Is it Community Detection, Anomaly Detection or Topology Classification ?
- **Within six months you should:**
  - Be able to compare the graph-based approach with traditional approach.
    - Can you combine both approaches in order to achieve better accuracy and lower false positive rate?

REFERENCES

- [1] A Word Graph Approach for Dictionary Detection and Extraction in DGA Domain Names: https://machine-learning-and-security.github.io/slides/Mayana-final-of-NIPS-DDGA.pdf
- [2] Oddball: Spotting anomalies in weighted graphs https://link.springer.com/10.1007%2F978-3-642-13672-6_40
- [3] Graph-based irregularity and fraud detection: http://www3.cs.stonybrook.edu/~leman/icdm12/ICDM12-Tutorial%20-%20PartI.pdf
- [4] Fast unfolding of communities in large networks: https://arxiv.org/pdf/0803.0476.pdf
- [5] Birds of a Feather Flock Together: The Accidental Community of Spammers https://www.cs.bgu.ac.il/~snean161/wiki.files/151_0605.pdf
- [6] Gephi: https://gephi.org
- [7] NetworkX: https://networkx.github.io

Think Graph!

Thank you!

Questions?

Mayana Pereira :: mpereira@infoblox.com