Process Query Systems for Cartel Detection

Vincent Berk
Dartmouth College
Hanover NH 03755 USA

vhb@Dartmouth.EDU
www.pqsnet.net

Background

Longstanding tradition in Law Enforcement technologies:

- 2001: EnCase digital image identifier scripts – used in 100s of court cases and settlements
- TOPOFF, TOPOFF2 participants
- Member of NYECTF
- NH and CT state computer forensics labs

Affiliates



Thayer School of Engineering at Dartmouth College

Tuck School of Business at Dartmouth

Dartmouth, Engineering and Business schools

ARDA

DTO

PROQSYS

US intelligence community

Commercial Software Production

Price Collusion

The United States Sentencing Commission estimates that 10% is the average overcharge resulting from an anti-trust offense, with actual instances as high as price doubling.

Ideal conditions are large numbers of customers, and a small number of firms (producers). Examples:

- Energy Market
- Pharmaceuticals
- Music Labels
- Oil and petroleum products

Understanding Detection

Price fixing agreements:	Made by “leaders” during a “communication”
Obfuscation techniques:	“Leaders” try to keep this “communication” secret
Detection strategy:	Analyze the social network structure for links or messages that are “out of place”

Social Network Analysis: The Enron Corporation

- Determine roles of individuals based on their observed behaviors.
- Funded by US-IC
- Dataset:
  - Email headers only
- After the fact analysis shows early warning would have been possible!



Social Network Background

Small World (Scale Free) Connectivity:

1. Growth: at each iteration a new node is connected to a randomly selected node in the existing network.
2. Preferential Attachment: the probability that the new node is attached to an arbitrary existing node *i* is directly proportional to the current connectivity of node *i*.

Scale-Free Properties of Random Networks (Barabasi, Albert, Jeong- 1999)

Limitation of small-world model:

1. No cost for maintaining relationships, so nodes may accrue an unbounded number of connections.
2. No consideration of relationship decay.
3. A better model for casual acquaintances or WWW page references than dynamic social networks.

A Referral Process for organization of dynamic social networks (properties):

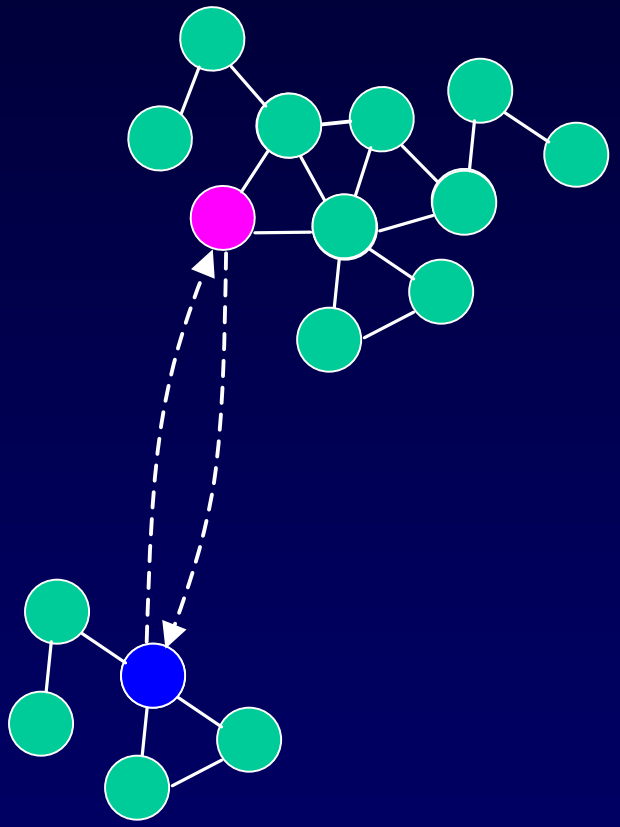
1. Limited degree: the probability of a person developing a new acquaintance falls off sharply as current number of friends reaches a certain level.

2. Clustering: the probability of two people becoming acquainted is significantly higher if they have one or more mutual friends.

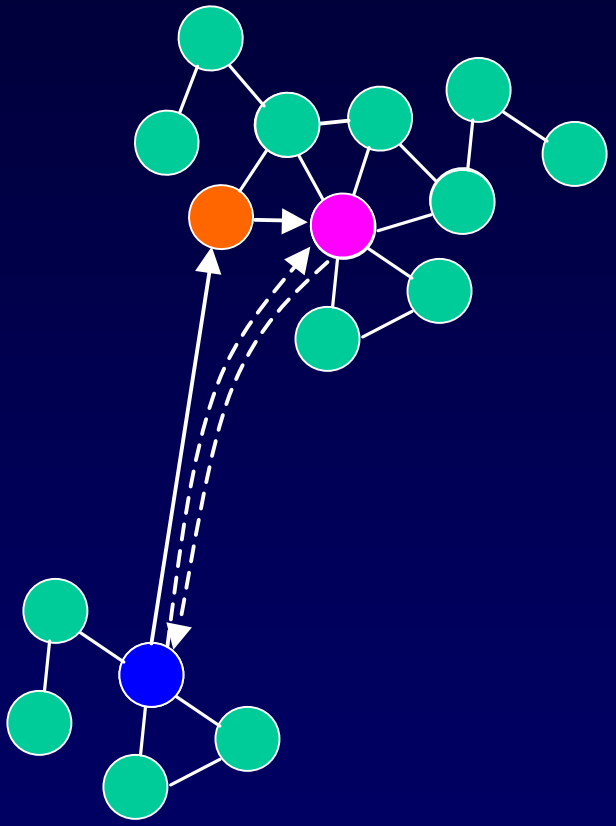
3. Decay of friendships: relationships may be broken with a probability or decay rate established by the strength of their mutual interests.

The structure of growing social networks (Jin, Girvan, Newman- 2001)

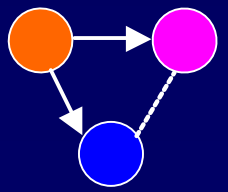
Subprocesses of Network Formation:



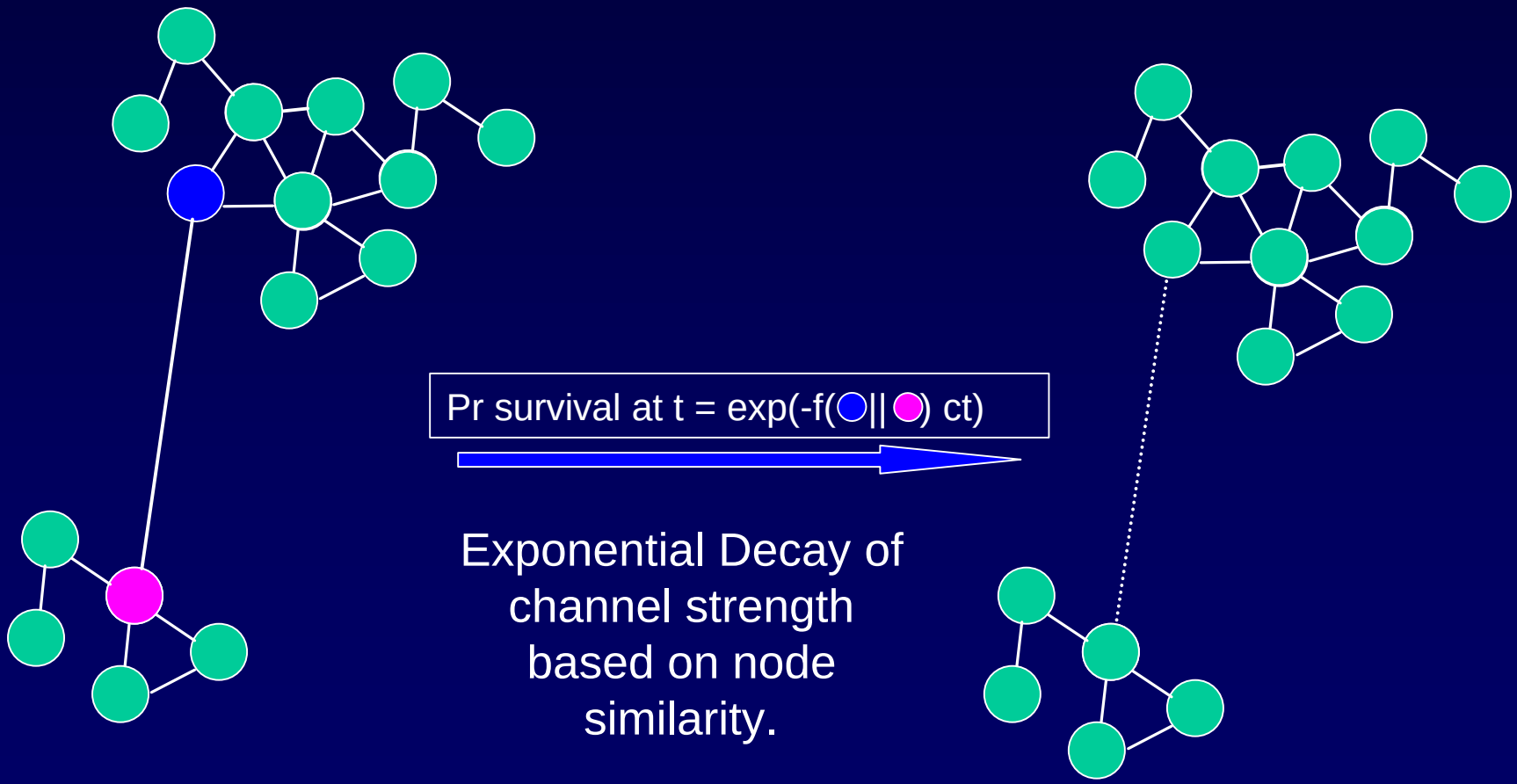
Via Random Encounter



Via Friendly Referral
(brokered by ●) :

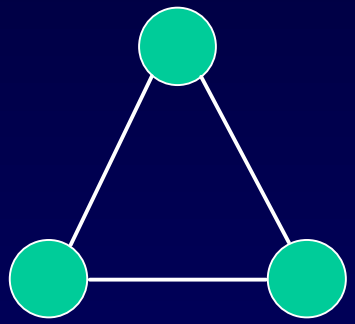


Subprocesses of Network Formation (Decay of Connections):

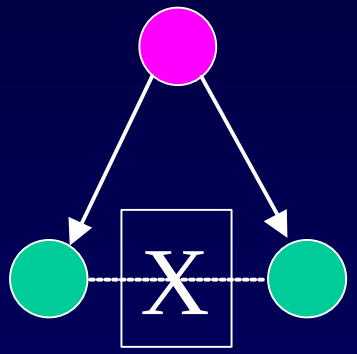


Exponential Decay of channel strength based on node similarity.

Structural Holes (Ron Burt 1995):



No potential for informational or economic advantage in fully connected relationship triad.



Structural hole allows ● to act as controlling agent in informational and economic transactions of the triad.

Messaging

Evidence of connections in social structures are observed messages:

- email, instant messages
- phone calls
- in person meetings

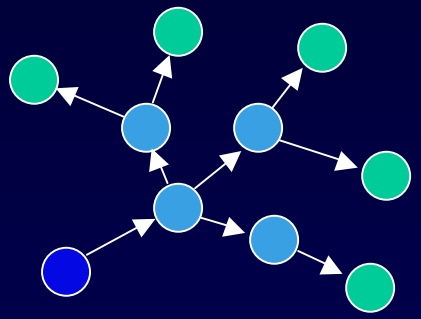
However: significant difference between message volume and importance of the social link

Instead: timing and sequence of messaging is more relevant

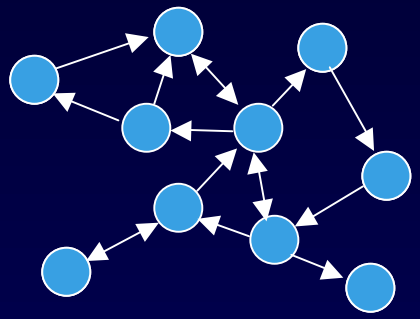
Additionally: link may exist while no messages are observed:

- there are none
- there are none now
- there are none in the observed space (eg. only email)

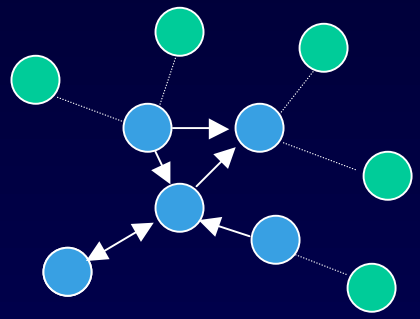
Thread Process Primitives:



Expanding
(coordinated or diffusive)

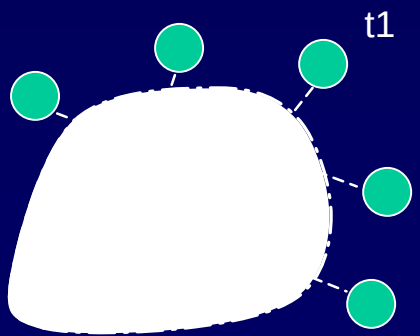


Stationary

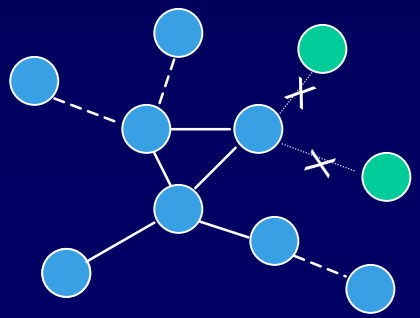


Contracting
(coordinated or diffusive)

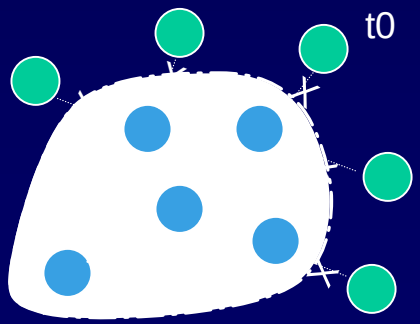
Network Formulation Process Primitives:



Expansion



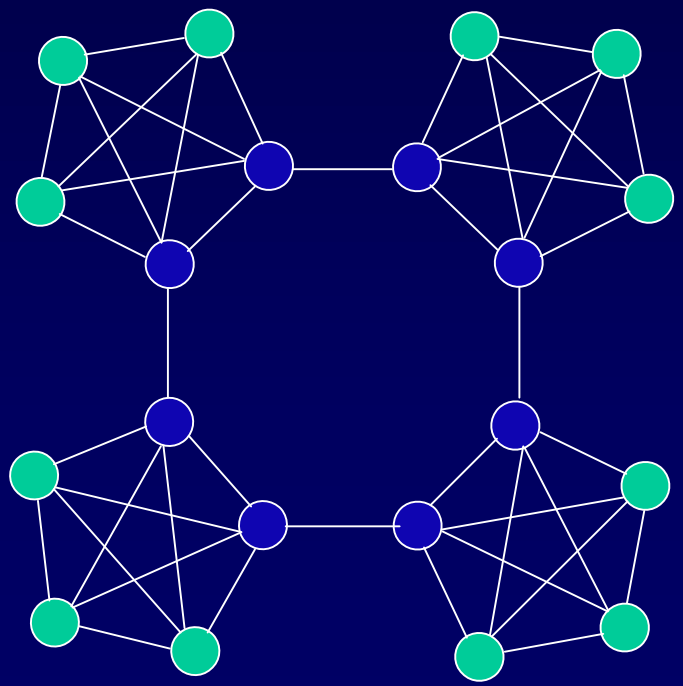
Consolidation



Contraction

Clustering (Leninist Networks)

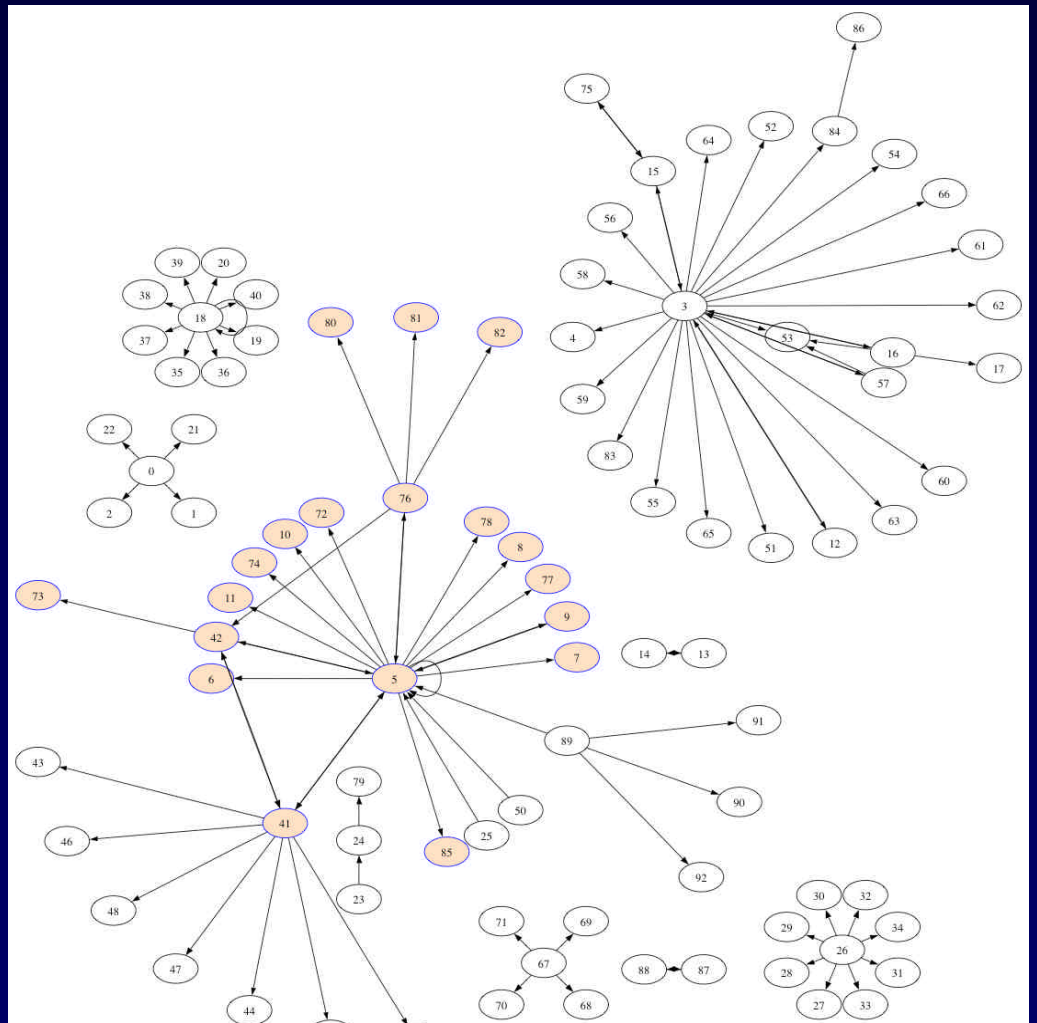
1. Isolated groups or *cells,* each with one or more leaders.
2. Inter-cell communication only through leaders.
3. Each cell leader only communicates with a small subset of other leaders.



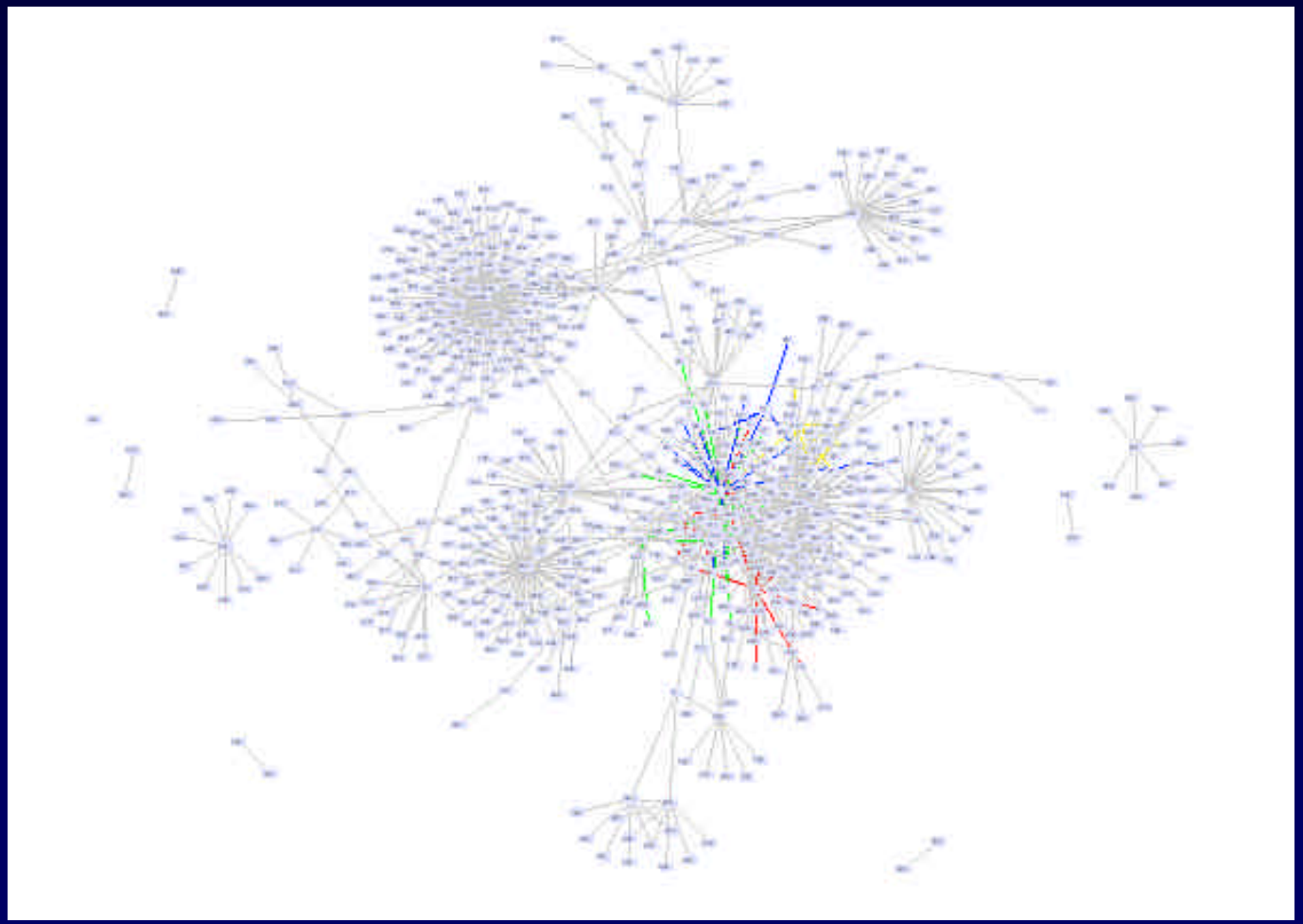
Leader node

(Coffman, Marcus- 2004)

Enron Data


86
75
64
52
84
54
15
66
56
61
58
3
62
4
53
16
17
57
59
83
60
55
65
51
12
63
39
20
38
40
18
37
19
35
36
80
81
82
22
21
0
2
1
76
72
78
10
8
74
77
73
11
9
42
14
13
7
6
5
91
89
43
79
50
90
41
24
25
85
46
92
48
23
71
69
67
30
32
29
34
88
87
26
47
28
31
70
68
27
33
44

Enron Data



Process Query Systems

Real-time detection of processes in observed event streams.

- Very resilient to noise, missed events
- Unlike rule-based systems, a PQS is scalable

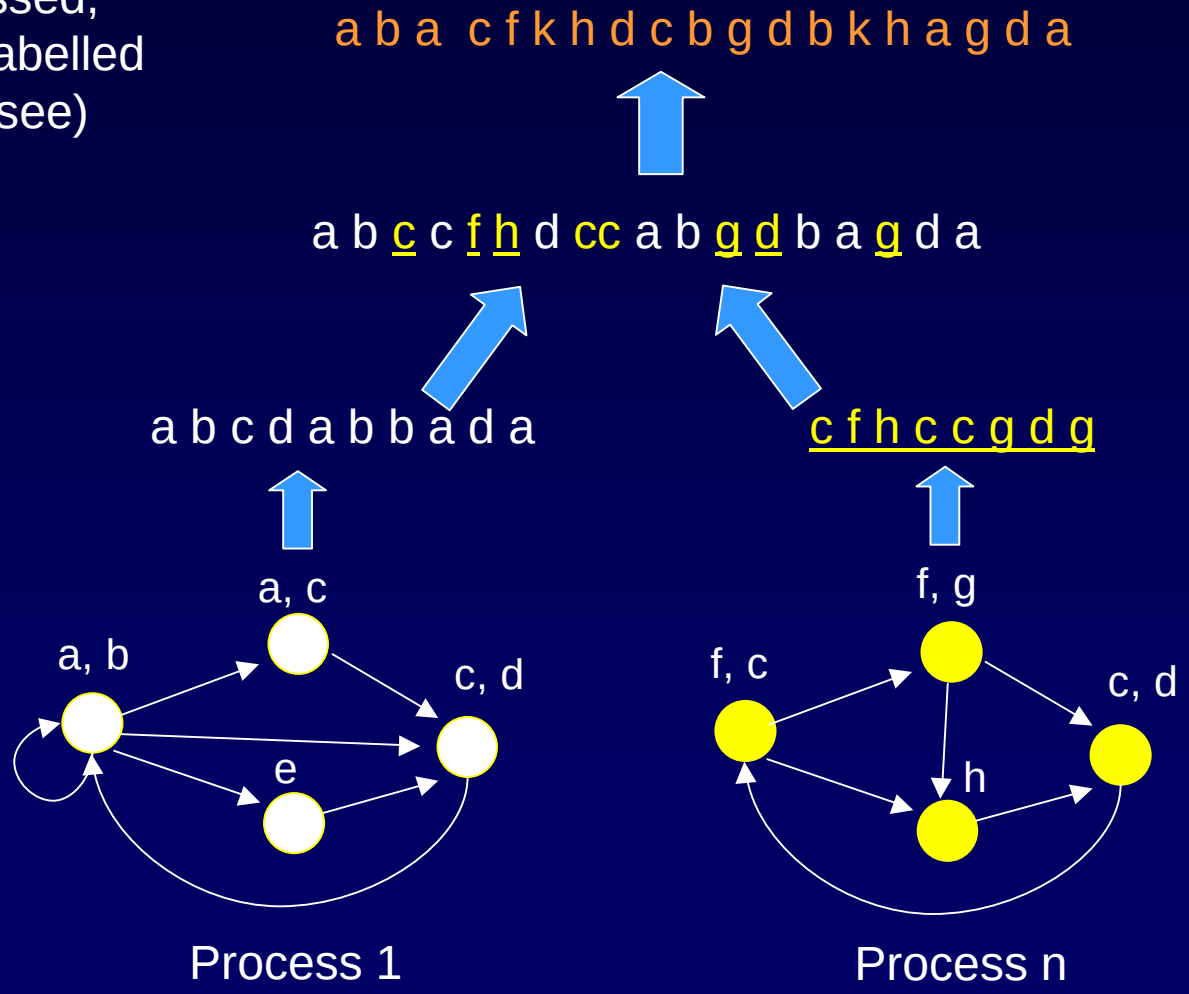
Problem Definition

Observations missed, noise added, unlabelled (This is what we see)

Observations are interleaved

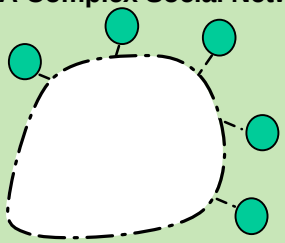
Observations related to state sequences

Underlying (hidden) state spaces



Basic Concepts of Process Query Systems (PQS)

A Complex Social Network



1 **consists of**

Multiple Processes

λ_1 = price fixing agreements

λ_2 = TPS reports

λ_3 = sales

2 **that produce**

Events



Real World

6 **that are used to do targeted inquiry**

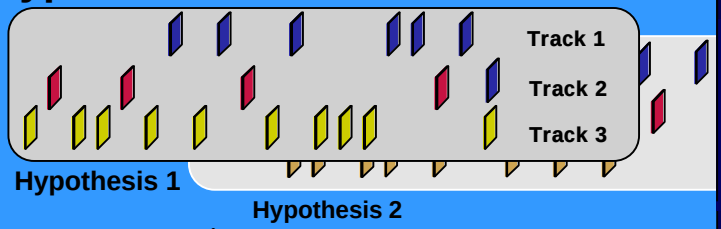
3 **that are seen as**

Indictors and Warnings

suspicious CEO meeting in London 3/17
CEO communicates directly to sales 3/19
.....

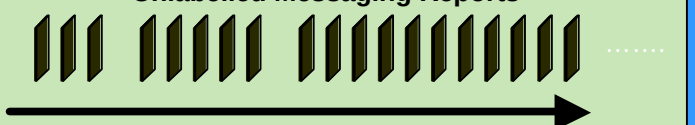
5 **that detect complex links and anticipate the next steps**

Hypotheses

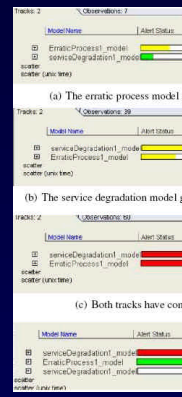


4 **that PQS resolves into**

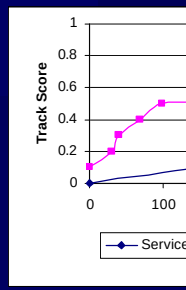
Unlabelled Messaging Reports



PQS



Sampl
Conso



Track
Score

Simple Collusion Detection

- Spatial-Temporal correlation of “leader” coordinates with changes in pricing structure.
- Inputs (mostly external):
  - Travel information
  - Corporate “leaders”
  - Pricing information
- Example: CEO “Amy” meets CEO “Bob”, both operating in the same market, to divide territories. Several weeks later, both firms change pricing around the same time.

Comprehensive Collusion Detection

- Based on controlling and “leader” nodes in a social network
- More predictive, identifies key people that must be involved to keep a fraudulent scheme hidden
- Inputs (mostly internal):
  - Email records, Instant Messaging
  - Phone records
  - Meeting notes
- Example: CEO meets only with CFO, and then emails all accounting personnel.

Summary

- Who is talking to who?
  - Meetings
  - Phone
  - Email
- If something is fishy, who would be the central person(s)?
  - Unconventional dissemination of information for a particular firm
- Information is necessarily incomplete and noisy, conventional methods do not scale.
- Privacy management!

Questions?

Vincent Berk
vhb@Dartmouth.EDU

Extra Slides

What is a PQS?

- New data processing paradigm
- Forms basis of many applications
- A Process Query System:
  - identifies,
  - tracks, and
  - predicts

  the behavior of objects in a given domain.

What can I do with it?

- Track fishes in an aquarium
- Detection and correlation of network security breaches, multistage attacks, and data exfiltration efforts.
- Tracking of targets by UAVs, scheduling their flight paths
- Identification and classification of vehicles and persons using a large sensor network
- Track and predict bird behavior at airports?

So what?

- Quick to apply to a domain

  observation handling, hypothesis generation, publishing/subscribing, and observation pairing is abstracted away from by the PQS

- Very resilient to noise, missed observations
- Unlike rule-based systems, a PQS is scalable

How?

- PQS is a generic engine
- Connect one or multiple data streams

  For instance: (x,y) coordinate pairs, IDS alerts, GPS coordinates, …

- Create a model describing dynamics of the environment

  For instance: maximum speed and turning radius of a vehicle (kinematic), order of events in a network attack (security)

Real time Fish Tracking



- Objective:

  Track the fish in the fishtank

- Why:

  Very strong example of the power of PQS

  – Fish swim very quickly and erratically
  – Lots of missed observations
  – Lots of noise
  – Kalman filters don’t work (non-linear movement and acceleration)

The fish

- 5 Gallon tank with 2 red Platys named Bill and Monica
- Camera generates a stream of “centroids”:
  For each frame a series of (X,Y) pairs is generated.
- Model describes the kinematics of a fish:
  The model evaluates if new (X,Y) pairs could belong to the same fish, based on measured position, momentum, and predicted next position. This way, multiple “tracks” are formed. One for each object.
- The whole system took only 3 days to build!

What will we see?

- Correct disambiguation of data stream:
  A tracked object leaves a 5 observation trail, *in the same color*.
- Correct tracking of fish:
  Colors are assigned arbitrarily, if a trail changes color often, we have trouble tracking.
- Correct prediction of path:
  When a fish swims behind an object and reappears, it's track is continued, concluding it is the same fish

Demo time







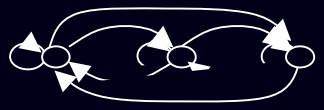

Improvements

- Higher resolution: the low resolution is unable to show the predictive abilities
- Movement characteristics of the fish: integrate actual fish behavior
- Sensor stream: add depth, size of fish, actual image processing techniques
- Tier-2 PQS is able to detect feeding behavior, chase behavior, and bubbles.

The Mathematical Problem Underlying PQS

Discrete Cause Separation Problem

Process Example:

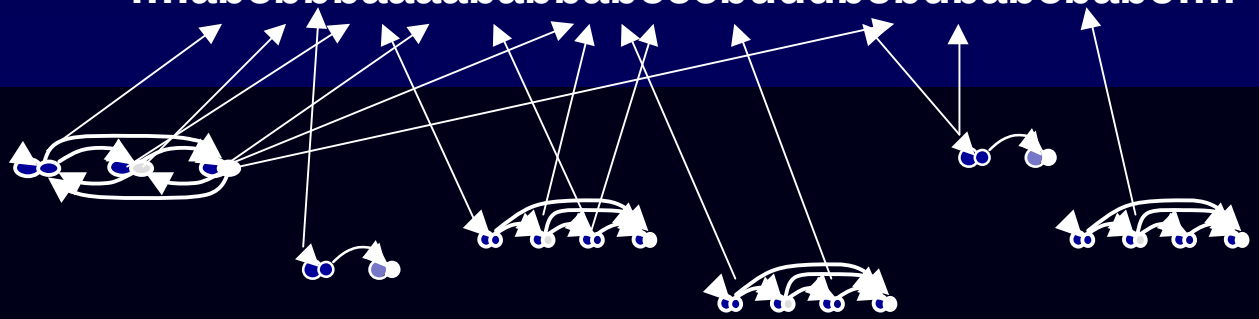


3 states + transition probabilities
n observable events: a,b,c,d,e,…
Pr(state | observable event) given/known

Observed event sequence:

….abcbbbaaaababbabcccbdddbebdbabcbabe….

Catalog of Processes



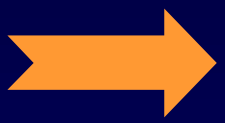
Which combination of which process models "best" accounts for the observations? This is what the PQS computes. Events not associated with a known process are "anomalies"

Revolutionary Potential of PQS

Databases

Pre-SQL

Domain Knowledge
+
File Management
+
Record Management



1970's

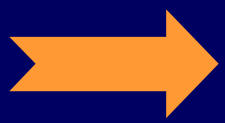
Post-SQL

Domain Knowledge

Process Detection

Pre-PQS

Domain Knowledge
+
Missed observation logic
+
Alert consistency



2000's ?

Post-PQS

Domain Knowledge

Summary

- PQS is a powerful new technology that identifies, tracks, and predicts the behavior of objects in a given domain.
- PQS is a generic core engine that abstracts away from data handling and hypothesis control, allowing the user to focus on building models.
- PQS models are domain specific and can often be built in under 8 hours.
- There is a fast and commercially supported version of the engine available.

The End

Questions?

For other projects based on PQS:

http://www.pqsnet.net

For further info:

vberk@Dartmouth.EDU

What does it mean to "detect" a process?

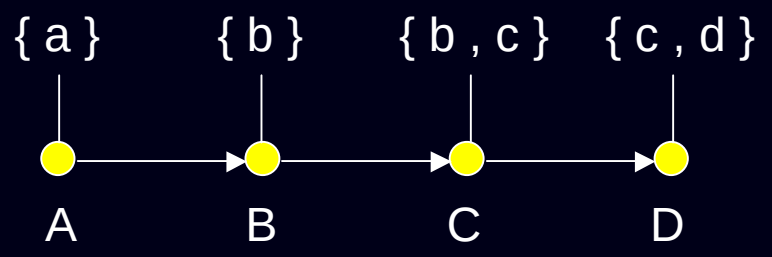
- Identify/estimate the process instance itself – eg network router failure or computer worm, truck or school bus, farmer or fedayeen
- Identify/estimate a process' state – eg how wide spread is the computer worm, what is the location of the truck, what activity is the fedayeen engaged in?
- Quantify the confidence in these estimates – eg. how good is the state estimate, how confident are we that a process model is the correct one, etc

With PQS, Users Focus on Models

- In practice, we build models of processes by:
  - First principles – ie, symmetry, physical laws, etc.
  - "Expert" models/rules/experience – ie, chess playing computers, military tactics, etc
  - Empirical analysis (from real or simulated data) – ie. backgammon, stock market models, etc.

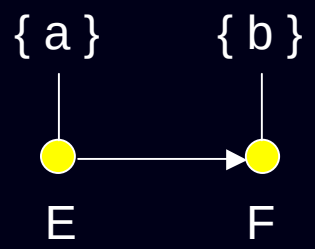
- Process Query Modeling Language was developed to allow for rapid insertion of new models into PQS

Detecting a Process Using Rules



WORM MODEL
(a,b,c,d ICMP traffic levels)

```
A,B,C,D = 0
repeat
        read event e
        if e==a then A
        if A and e==b then B
        if B and (e==b or e==c) then C
        if C and (e==c or e==d) then D
until D
```

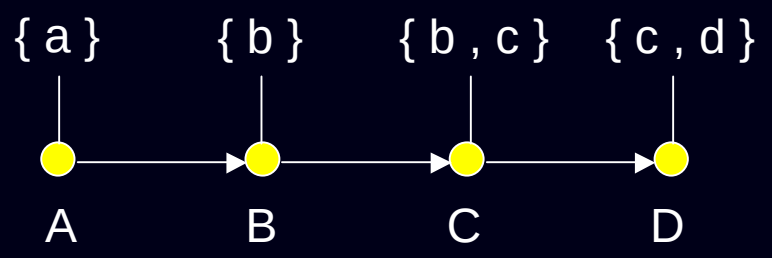


ROUTER FAILURE MODEL

```
E,F = 0
repeat
        read event e
        if e==a then E
        if E and e==b then F
until F
```

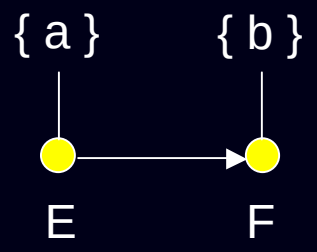
What does “ab” mean ? (Process ambiguity)
What does “ac” mean ? (Missed Detections)

Now Add Rules for Process Disambiguation



WORM MODEL
(a,b,c,d ICMP traffic levels)

```
A,B,C,D = 0
repeat
        read event e
        if e==a then A
        if A and e==b then B
        if B and (e==b or e==c) then C
        if C then (E=0, F=0)
        if C and (e==c or e==d) then D
until D
```

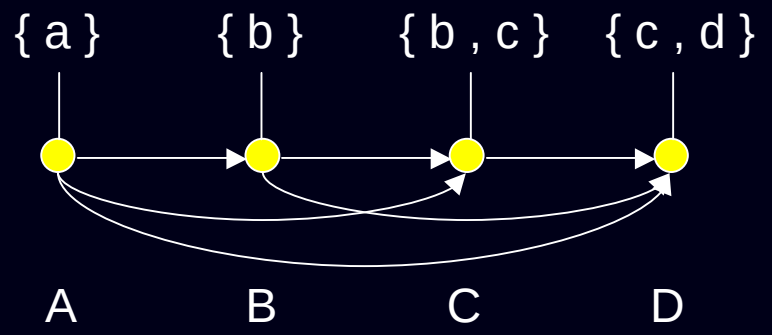


ROUTER FAILURE MODEL

```
E,F = 0
repeat
        read event e
        if e==a then E
        if E and e==b then F
until F
```

Cannot decide which process is present until more data arrives

Now Add More Rules for Missed Detections, etc, etc...

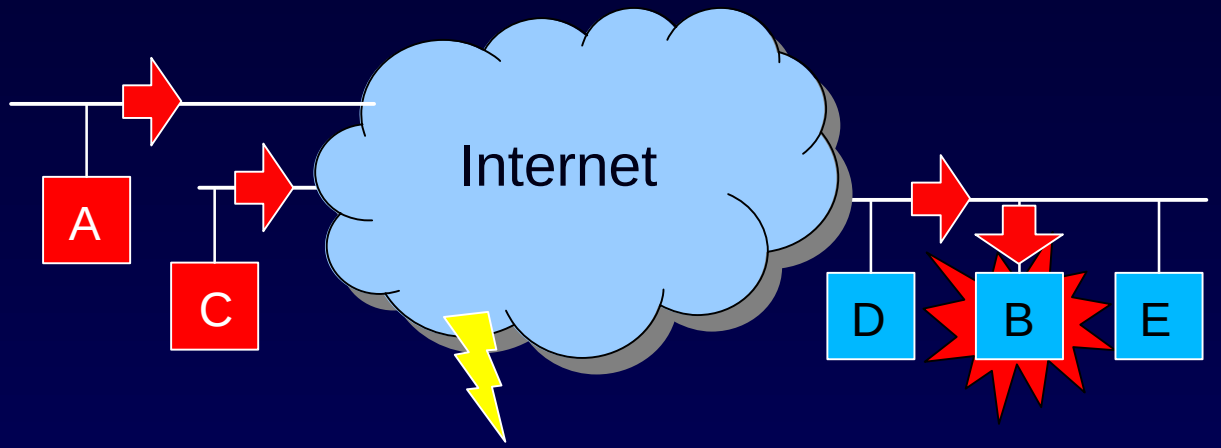


WORM MODEL
(a,b,c,d ICMP traffic levels)

```
A,B,C,D = 0
repeat
        read event e
        if e==a then A
        if A and e==b then B
        if A and e==c then C,D
        if A and e==d then D
        if B and (e==b or e==c) then C
        if C then (E=0, F=0)
        if C and (e==c or e==d) then D
        if D then (E=0, F=0)
until D
```

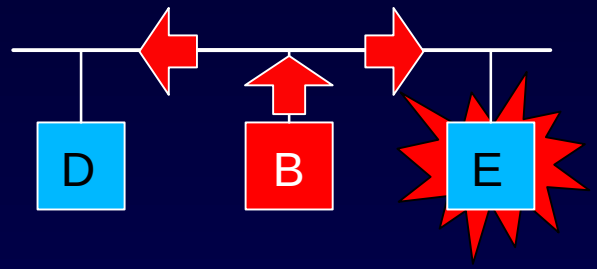
This clearly does not scale and does not lead to manageable sets/systems of rules

Example Scenario



Tier1 Alerts	Indicators
A scans B	Snort: 02/21-20:06:17.904500 [**] [1:469:1] ICMP PING NMAP [**] [Classification: Attempted Information Leak] [Priority: 2] {ICMP} 131.58.63.160 -> 100.10.20.4
C attacks B (success)	SSL error log (host 100.10.20.4): [Mon Feb 21 20:30:24 2005] [error] mod_ssl: SSL handshake failed (server www.osis.gov:443, client 138.185.170.45) (OpenSSL library error follows) [Mon Feb 21 20:30:24 2005] [error] OpenSSL: error:1406908F:lib(20):func(105):reason(143)

Example Cont'd



Tier1 Alerts	Indicators
B scans D	02/21-20:31:17.528602 [**] [1:1807:2] WEB-MISC Chunked-Encoding transfer attempt [**] [Classification: Web Application Attack] [Priority: 1] {TCP} 100.10.20.4:34074 -> 100.10.20.169:80
B attacks D (fails)	100.20.1.169 - - [21/Feb/2005:08:31:22 -0500] "GET /default.idq?AAAAAAAAAAA……..AAAAAAA HTTP/1.1" 404 1287 "-" "-"
B scans E	02/21-20:32:01.622465 [**] [1:1807:2] WEB-MISC Chunked-Encoding transfer attempt [**] [Classification: Web Application Attack] [Priority: 1] {TCP} 100.10.20.4:34076 -> 100.10.20.170:80
B attacks E (succeeds)	100.20.1.170 - - [21/Feb/2005:08:32:06 -0500] "GET /default.idq?AAAAAAAAAAA……..AAAAAAA HTTP/1.1" 200 1287 "-" "-"