SAND2013-4550C

Information-Theoretic Measures for Security Informatics

Rich Colbaugh **Travis Bauer** **Kristin Glass**

Sandia National Laboratories

December 2012

Introduction

Objective

Explore extent to which concepts and measures from information theory can be leveraged to solve challenging security informatics problems.

Outline

- Some information theory:

  comparing distributions, stochastic dynamics/Markov chains (MC), comparing stochastic dynamics – Kullback-Leibler divergence rate.

- Understanding and predicting social network dynamics:

  transfer entropy, "influential" individuals, predicting the activities of individuals.

- Dynamics-based malware detection:

  Markov chain (MC) models for executables, malware classification based upon MC similarity.

Some Information Theory

Comparing distributions

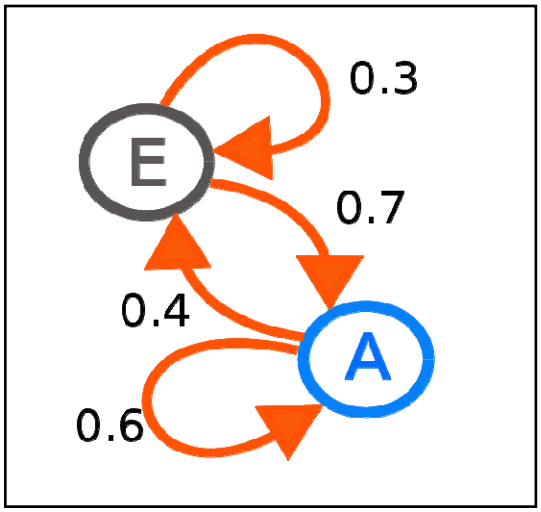
- *Entropy* $H_I = -\Sigma$ p(i) log p(i) measures average number of bits needed to optimally encode independent draws of discrete random variable I. This can be interpreted as uncertainty/information in distribution p(i).
- *Kullback-Leibler divergence* $KL_I = \Sigma$ p(i) log(p(i)/q(i)) measures excess number of bits (over optimal) needed to encode I if distribution q(i) is used in place of p(i). This can be interpreted as measure of difference between distributions p(i) and q(i).
- *Mutual information* of random variables I,J with distribution p(i,j), $MI_{IJ} = \Sigma$ p(i,j) log (p(i,j)/p(i)p(j)), is excess code needed if erroneously assume I,J are independent (this follows directly from KL_I). This symmetric measure can be interpreted as extent to which knowing I reduces uncertainty concerning J.

Some Information Theory

Stochastic dynamics

The preceding measures are most naturally interpreted in terms of "static" distributions. To generalize to stochastic dynamics we consider a simple but useful class of dynamical systems: Markov chains.

- A *Markov chain* MC = {Q, P} consists of a state set $Q = \{q_1, \ldots, q_n\}$ and a matrix of transition probabilities $P = [P_{ij}]$ which characterize Markov dynamics: $\Pr(q_{t+1} = j \mid q_t = i, q_{t-1} = k, \ldots) = \Pr(q_{t+1} = j \mid q_t = i) = P_{ij}$.
- A *Markov chain of order k* depends upon additional history: $\Pr(q_{t+1} \mid q_t^{(k+1)}) = \Pr(q_{t+1} \mid q_t^{(k)}) = \Pr(q_{t+1} \mid q_t, q_{t-1}, \ldots, q_{t-k+1})$.
- It is frequently convenient to represent a MC as a weighted directed graph.



Some Information Theory

Comparing stochastic dynamical systems

Here we derive the Kullback-Leibler divergence rate as a measure of the difference between two (stationary) MC.

- *Entropy rate* $h_I = -\Sigma\ p(q_{t+1}, q_t^{(k)})\ \log\ p(q_{t+1} \mid q_t^{(k)})$ (for stationary MC) measures average number of bits needed to encode one additional state in sequence defining dynamics. This can be interpreted as the growth rate of the dynamics' uncertainty or information.
- *Kullback-Leibler divergence rate* $kl_I = \Sigma\ p(q_{t+1}, q_t^{(k)})\ \log\ [p(q_{t+1} \mid q_t^{(k)}) / p(s_{t+1} \mid s_t^{(k)})]$ measures excess number of bits needed for coding if MC {S, R} is used in place of MC {Q, P}. This can be interpreted as a measure of the difference between {Q, P} and {S, R}. If $k = 1$ and {Q, P} has stationary distribution π then $kl_I = \Sigma\ \pi_i\ P_{ij}\ \log\ (P_{ij} / R_{ij})$.
- Note: kl_I is non-parametric in that it involves no assumptions about the distributions (e.g., Gaussian) or model structure (e.g., linearity).

Social Network Dynamics

Problem formulation

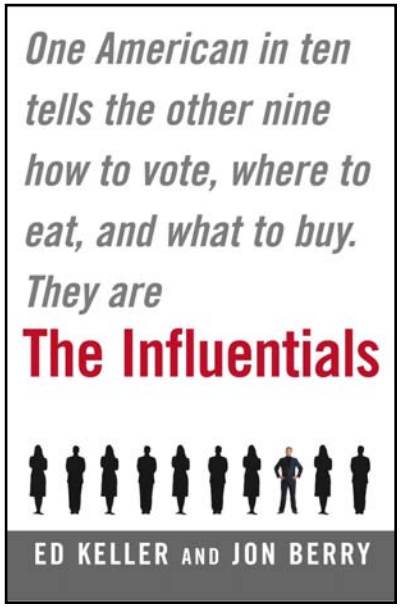
- Objective: given observations of activities of a population of interest, 1.) identify "influential" individuals in the population, 2.) predict the future behavior of individuals.
- Formulation:
  - use observations of individual activities to infer "meaningful" social network $G = (V,E)$, in which edge $e \in E$ reflects an influence or predictive relationship between agents;
  - leverage G to identify influentials/predict activities.

Do I need to remind you that I have a huge Twitter following?

Social Network Dynamics

Transfer entropy

- Hypothesis: we can characterize pathways of influence and assess the predictability of agent activity in a social network by quantifying, for each agent pair (I,J), the extent to which the behavior of J predicts or influences that of I.
- Given measurements of the activities of I and J over time, say $\{i_t, i_{t-1}, i_{t-2}, \ldots\}$ and $\{j_t, j_{t-1}, j_{t-2}, \ldots\}$, it is natural to model J's influence on I via an *influence MC,* $MC_I = \{Q, P, S\}$, where S is J's state set and P is a tensor of transition probabilities: $P_{ijk} = \Pr(q_{t+1} = j \mid q_t = i, s_t = k)$.
- This is easily extended to *higher-order influence MC,* for which $P_{ijk} = \Pr(q_{t+1} \mid q_t^{(m)}, s_t^{(n)})$.



Social Network Dynamics

Transfer entropy (cont'd)

The Kullback-Leibler divergence rate kl_I provides a measure of the excess code needed if it is erroneously assumed that I's behavior is independent of J – this is the *transfer entropy* (TE) from J to I:

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})} .$$

TE thus measures the extent to which knowledge of J's activity reduces uncertainty concerning I's activity. Consequently TE possesses predictive power (large $TE_{J\to I}$ implies J's behavior predicts I's behavior) and gives a sense of influence (TE is a nonlinear extension of Granger causality).

Determining whether an estimated TE is significant is challenging; we compare estimated TE with TE obtained when J's time series is shuffled.

Social Network Dynamics

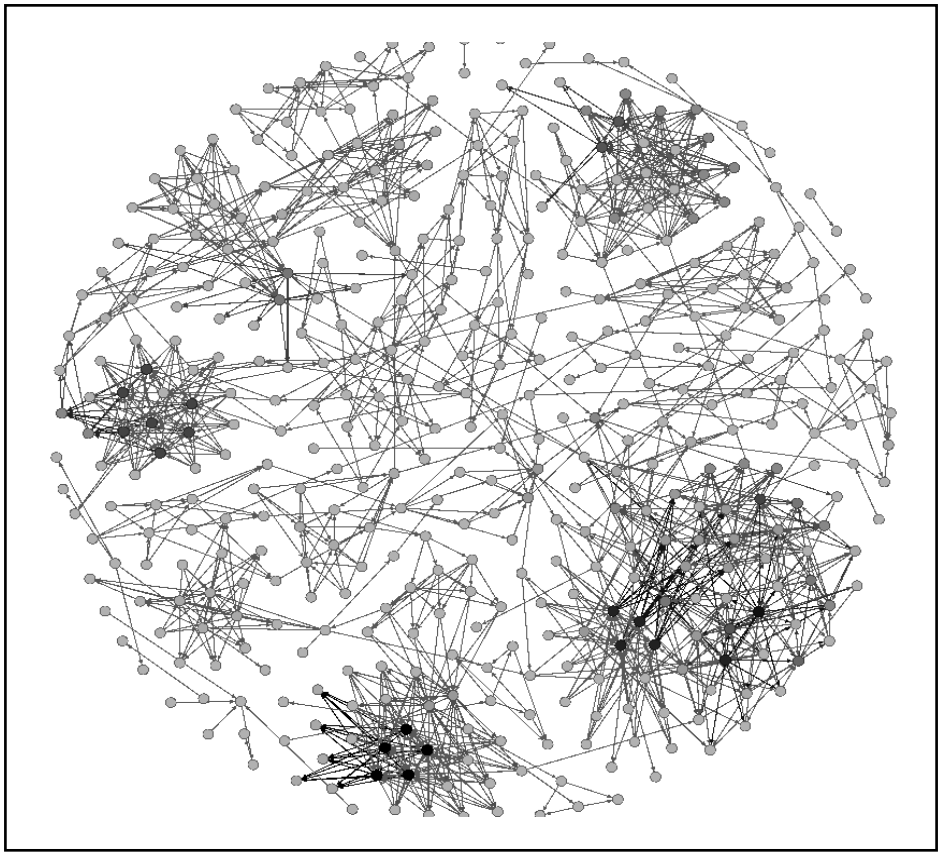
Identifying influential individuals

- Phenomenon: influential editors in Wikipedia (WP).
- Approach:
  - construct (weighted, directed) TE graph from editing time series;
  - define J to be "influential" if J's editing behavior predicts that of several other editors – $TE_{J \to I}$ is large for several I – and identify influentials by finding editors with large out-degree in TE graph.
- Sample results:
  - for WP 'anarchism' page, five editors (of ~1200) account for ~27% (54/201) of tails for top TE edges (weekly series, memory = 3; similar for daily series, memory = 7);
  - for WP 'Elvis' page, six editors (of ~1200) account for ~28% (50/179) of tails for top TE edges (daily series, memory = 7).

Social Network Dynamics

Identifying influential individuals (cont'd)

Sample visualization: TE graph for 'anarchism' page (top 1300 edges).



Social Network Dynamics

Predicting activity of individuals

- Phenomenon: editing behavior in WP.
- Approach: given time intervals of interest $[0,T_1]$, $(T_1,T_2]$
  - construct TE graph from editing activity during $[0,T_1]$;
  - for individual of interest I, predict activity on $(T_1,T_2]$ based upon 1.) I's activity on $[0,T_1]$, 2.) I's TE neighbors' activity on $(T_1,T_2]$.
- Sample results: for WP 'anarchism' page and the task of predicting whether a given (active) editor edits during the upcoming week,
  - we obtain an average prediction accuracy of 81.6% (compared with 60% for random guessing);
  - considering TE graph boosts prediction accuracy from 77.1% (history-only) to 81.6%, indicating methodology has potential.

Malware Detection

Problem formulation

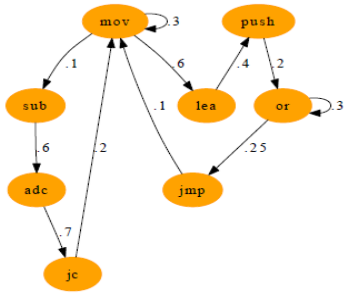
- Objective: improve capability to detect/counter advanced cyber threats by considering *dynamical* features of attacker activity.
- Preliminary study: malware detection.
- Formulation:
  - model time series of (dynamical) instruction traces of executables as MC;
  - characterize each executable/MC in terms of its KL divergence-rate "distance" to all other executables/MC in dataset;
  - learn classifier (e.g., bipartite graph RLS) which accurately predicts class {malware, legitimate} of given executable using (KL-based) MC distance profile as feature vector.

Malware Detection

Approach

- Data: time series for (dynamically collected) instruction traces for 780 malware instances and 776 benign programs [Anderson et al. 2012], modeled as MC.

mov	esi, 0x0040C000
lea	edi, [esi-0xB000]
push	edi
or	ebp, 0xFF
or	ebp, 0xFF
jmp	0x00419532
mov	ebx, [esi]
mov	ebx, [esi]
sub	esi, 0xFC
adc	ebx, ebx
jc	0x00419528
...	...



- Instance features: each instance I of malware/goodware is represented by the vector $\mathbf{x}_I = [x_{I1}, \ldots x_{IN}]^T$ of distances to all other programs, with distance x_{IJ} between $MC_I = \{Q,P\}$ and $MC_J = \{S,R\}$ quantified by the "symmetrized" KL divergence rate

$$kl_S = kl(P,R) + kl(R,P)$$
$$= \Sigma\, \pi_i\, P_{ij} \log (P_{ij} / R_{ij}) + \Sigma\, \theta_i\, R_{ij} \log (R_{ij} / P_{ij}).$$

- Classification: we learn classifier $f(\mathbf{x}_I) \rightarrow$ {malware, goodware} using our bipartite graph RLS method [Colbaugh/Glass 2011].

Malware Detection

Preliminary study

- We test our proposed malware detection method using ten-fold cross-validation with the 1556 malware/goodware programs, and compare the performance of this information theory-inspired approach with that of two other methods: 1.) classifier using n-grams of instruction traces, 2.) commercial, signature-based antivirus programs.
- Results: accuracy of methods
  - proposed method = 89.5%;
  - n-gram method = 80.6%;
  - commercial AV = 75.3%.

the ONION
America's Finest News Source

NATIONAL

48-Hour Internet Outage Plunges Nation Into Productivity

OCTOBER 1, 2003 | ISSUE 39-38

BOSTON—An Internet worm that disabled networks across the U.S. Monday and Tuesday temporarily thrust the nation into its most severe maelstrom of productivity since 1992.

ARTICLE TOOLS

DIGG