Influence Maximization on Complex Networks with Intrinsic Nodal Activation

Methods and Applications

Presenter: Arun Sathanur[1]

Collaborators: Mahantesh Halappanavar[1], Yi Shi[2] and Yalin Sagduyu[2]

[1] Pacific Northwest National Labs, Richland, WA

[2] Intelligent Automation Inc., MD

Outline

A brief introduction to Influence Maximization

Intrinsic and Influenced Activation

Submodularity of the Influence Function

Experimental Results

An Approximate Linear Model for Activation Probabilities

Attack modeling in Cyber Systems

Conclusions and Future Work

Influence Maximization Problem - A Brief Introduction

General Influence Maximization Problem

- A large class of natural and man-made systems with rich dynamics can be studied through the abstraction of graphs.
- Signals arrive at each node along the incoming edges, undergo (non-linear) processing at the node and the processed signal is transmitted along the out-going edges.
- Given the models for node behavior and edge interactions and the objectives we are interested in, how can we find those influential nodes which have maximal impact on the system ?
- Applications in diverse fields
  - Viral marketing for product adoption
  - Spread of content on social media
  - Spread of diseases in contact-networks
  - Keystone species in microbial communities
  - Controllability and Observability in complex systems

The influence maximization problem

- Given: A graph ***G(V,E,ω)***, a diffusion model (how a vertex gets activated based on the state of its neighbors), and a budget **k**, the influence maximization problem is stated as follows:

- Find a set of **k** vertices called the seed set **S**, that when activated results in maximal activations on the network amongst all possible sets of **k** vertices

- Two most popular diffusion models
  - Linear Threshold: A vertex can get activated if a fraction of neighboring vertices that are active is greater than a threshold Θ_v
  - Independent Cascade: One shot chance for an activated vertex to activate its neighbor

2003: Kempe, Kleinberg, Tardos.

Modes of Activation

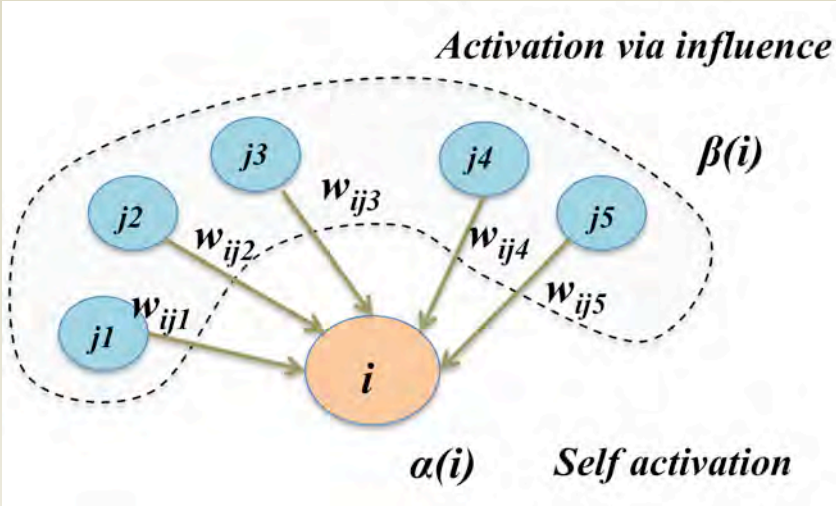
Intrinsic activation vs Influenced Activation

Activation at each node is split into two mechanisms to allow correspondences to real-world situations

Intrinsic activation: Activation at each node attributable to it's own intrinsic mechanisms

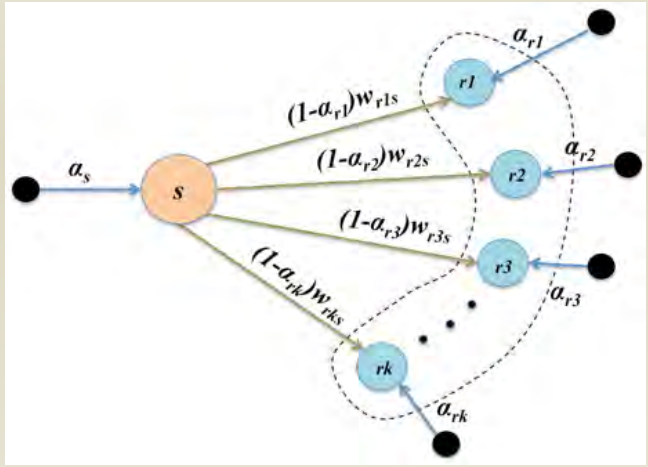
Influenced activation: Activation originating at each node attributable to influence of the neighboring nodes

Parameterize by the tendency towards intrinsic activation denoted by α and that towards influenced activation by β



Arun V. Sathanur and Mahantesh Halappanavar, "Influence Maximization on Complex Networks with Intrinsic Nodal Activation," Social Informatics, Bellevue, WA, Nov 2016

Content Creation vs Consumption / Spread



$$\sigma(s) = \alpha_s \sum_{k=1}^{d_s^o} (1 - \alpha_{r_k}) w_{r_k s}$$

Activation via influence creates engagement on the social network platform

All *α* values as high => Nodes create high volume of content on their own
Not much spreading of the content through engagement.

All nodes having small values of α => Nodes are eager to spread the content but there is not much content created in the first place, again reducing the engagement.

Hypothesis: There is an optimal assignment of the *α* values for a given network topology that can maximize the spread of influence under intrinsic activation.

Mining Influential Nodes under the IC Model with Intrinsic Activation

Algorithm 1 Selects a set of k influential nodes that cause maximal activations on a network, following the independent cascade (IC) model with self-activation (IC-Int). The inputs are a directed graph ($G = (V, E)$), set of edge probabilities ($P = \{p_{uv} : (uv) \in E\}$), vector of alpha values ($\alpha = \{\alpha_v : v \in V\}$), number of samples ($n$), and number of influential nodes to be identified (k).

procedure IC-INT(G, P, α, k, n)
  Generate n random numbers $r_{uv}^1 \ldots r_{uv}^n$ for each edge in E and generate a set SG containing n subgraphs such that in subgraph i, $p_{uv} \geq r_{uv}^i$
  $S \leftarrow \emptyset$ ▷ Set of influential nodes to be mined
  while $|S| < k$ **do**
    $v_{best} \leftarrow \emptyset$, $a_{best} \leftarrow 0$
    for each node v in $V \setminus S$ **do**
      $a \leftarrow 0$
      for each $G_i \in SG$ in **parallel do**
        $\hat{S} \leftarrow$ active nodes in $S \cup \{v\}$ based on α
        Compute number of nodes, $\hat{a}$, in $V \setminus \hat{S}$ that are reachable from the $\hat{S}$
        $a \leftarrow a + \hat{a}$ ▷ Synchronized update
      if $a \geq a_{best}$ **then**
        $v_{best} \leftarrow v$
        $a_{best} \leftarrow a$
    if $v_{best} \neq \emptyset$ **then**
      $S \leftarrow S \cup \{v_{best}\}$
  return S

Submodularity of the Influence Function

Influence Maximization as Submodular Optimization

Consider a set of entities *V*. Let $S \subseteq V$. Let $f : 2^V \rightarrow \mathcal{R}$, be a mapping that associates every such *S* with a real number. f is submodular provided it satisfies the following.

For every $S \subseteq T \subseteq V$ and any $v \in V, v \notin T$ we have

$$\boldsymbol{f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)}$$

A function $f : 2^V \rightarrow \mathcal{R}$ is monotone provided, for any $S \subseteq T$, $f(S) \leq f(T)$

We want to select the optimal subset called the seed set S such that $f(S)$ is maximized under a constraint on the cardinality of the seed set *S i.e* $|S| \leq k$.

For this problem the simple Greedy algorithm works very well

Let S^* denote the optimal solution to the problem. If *f* is monotone submodular and $f(\phi) = 0$, then due to a powerful result derived by Nemhauser et al.

$$\boldsymbol{f(S) \geq \left(1 - \frac{1}{e}\right) f(S^*)}$$

Sub-modular property under intrinsic activation

Total number of activations on the network is submodular for IC (KKT-03)

Total number of activations on the network is ***NOT*** submodular for IC-INT

However total number of activations is not a good influence metric for IC-INT

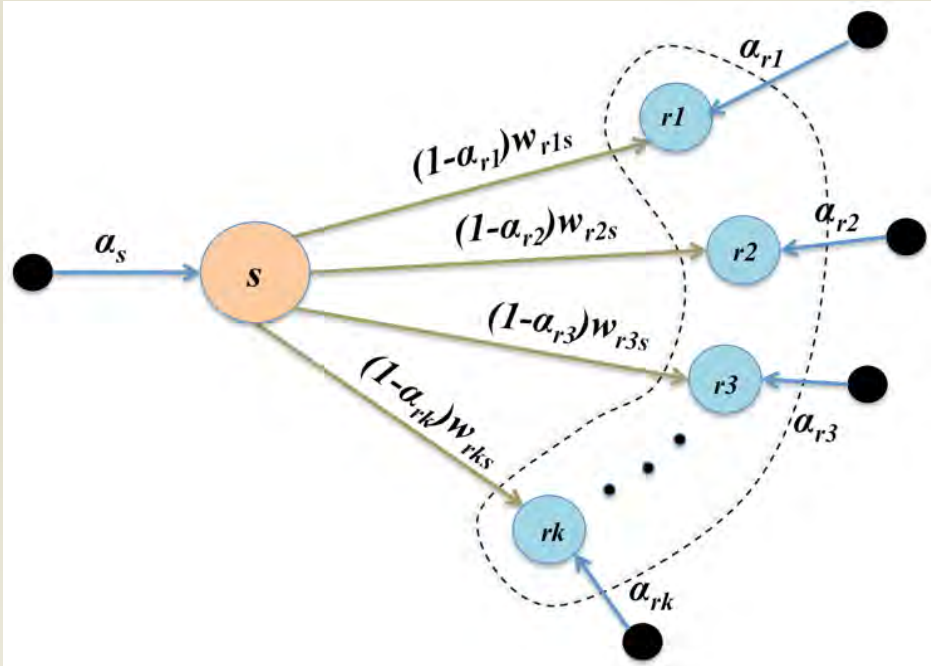
Better metric would be the total number of activations attributable to influence

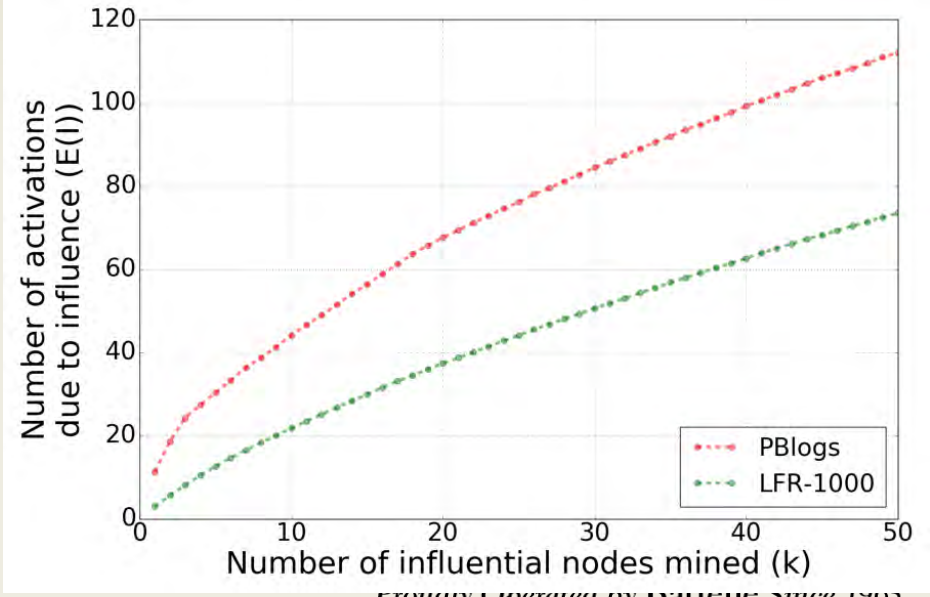
Introduce one dummy node and a dummy directed edge for each node on the network. IC probability for the dummy edge is the node α value

Over the set of dummy nodes we have an IM problem with IC model. The modified influence metric is then submodular

Evidence of SubModularity

PBlogs and LFR Networks





Experiments on a Twitter Graph

Twitter Dataset

The Twitter network around Purdue Engineering students

Started with a seed user
Expanded by doing a BFS (Collecting Followers)
Filtered low degree nodes

Eventual graph had 1167 nodes and 10292 edges

Estimated the ***α*** and ***W*** parameters via ML estimates by counting the various types of tweets

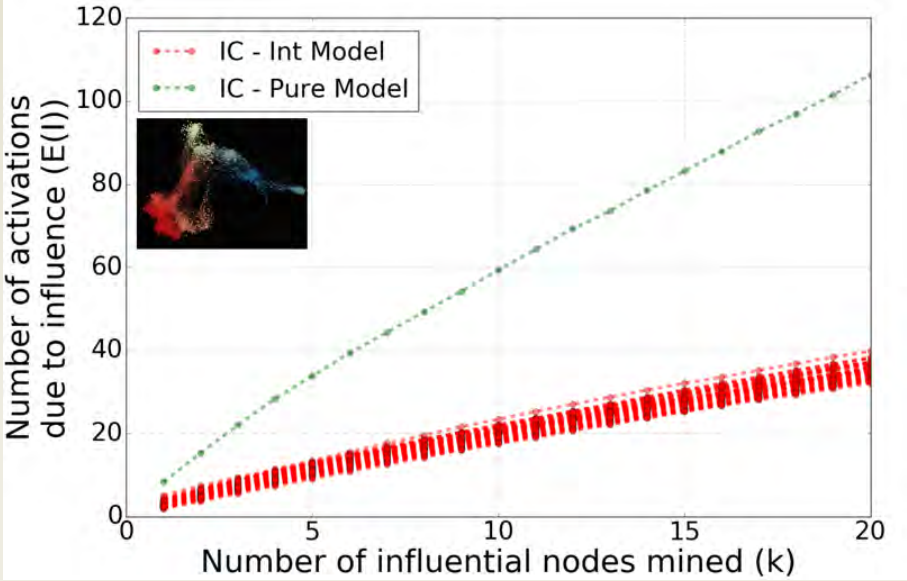
$$\gamma_i = \sum_j \gamma_{(j,i)},$$
$$\alpha_i = \frac{k_i}{\gamma_i + k_i},$$
$$\beta_i = 1 - \alpha_i,$$
$$w_{ij} = \frac{\gamma_{(j,i)}}{\gamma_i}.$$

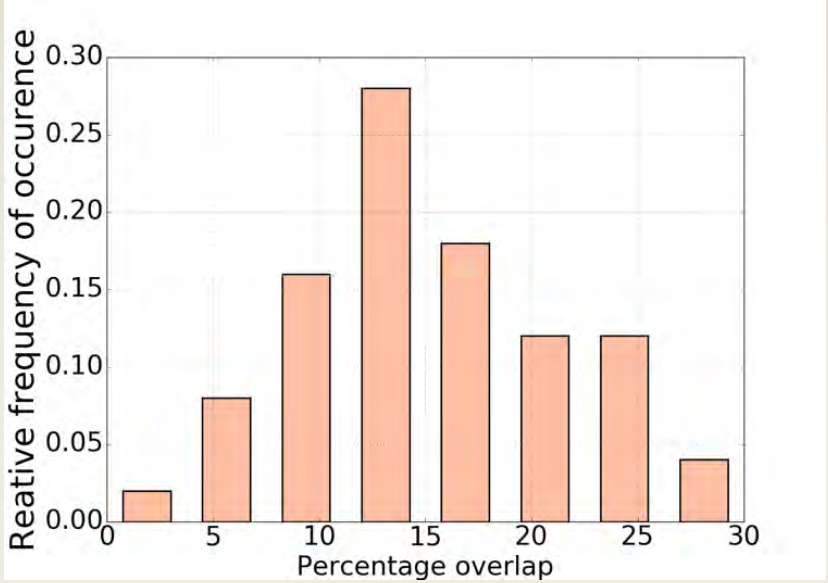
k_i = Number of intrinsic tweets from user *i*
γ_i = Total number of interactions that user *i* participated in.
$\gamma_{(j,i)}$ = Number of interactions interactions of user *i* with user *j*

Experiments on the Twitter network

Independent Cascade model over-estimates the number of activations

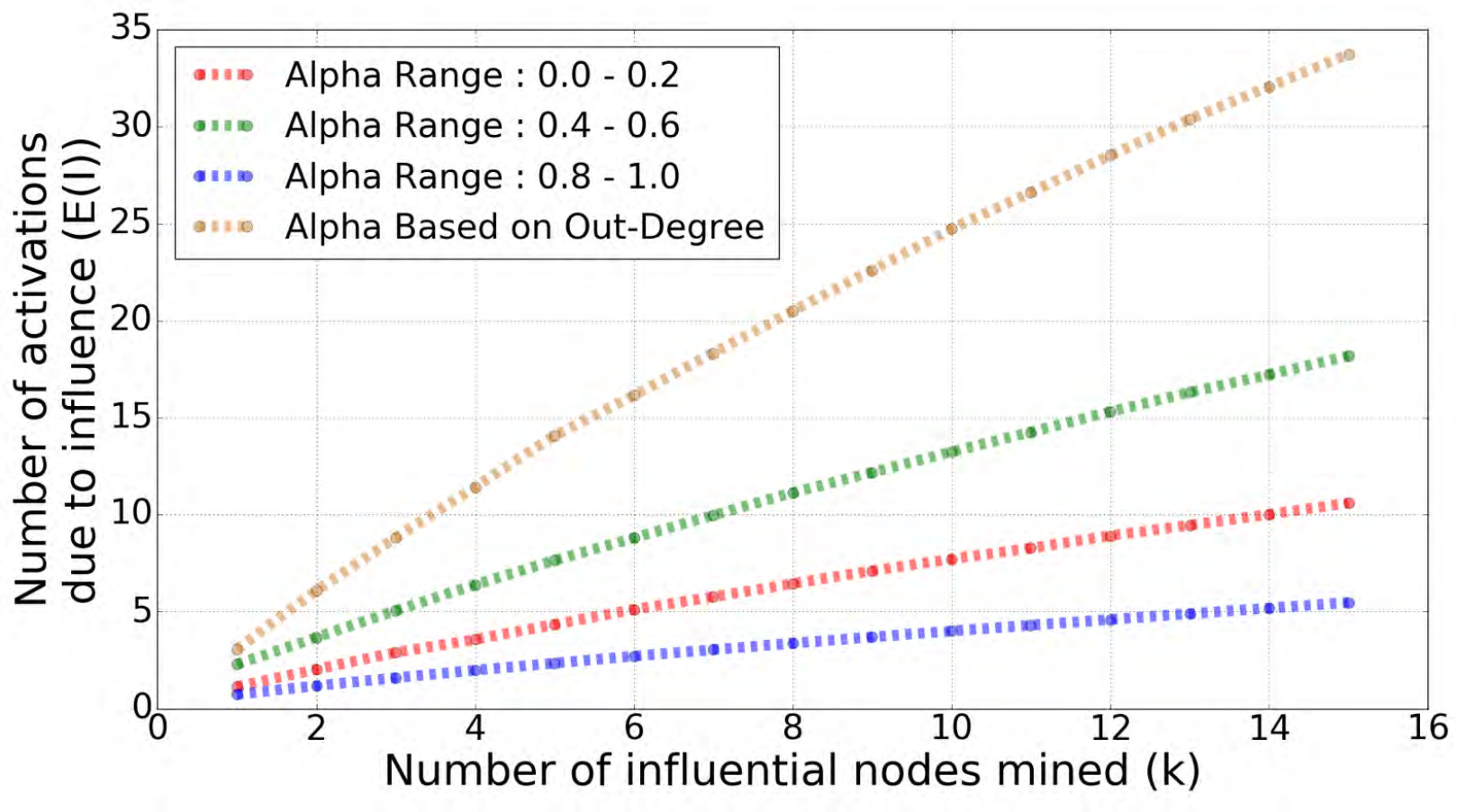
A small but significant fraction of the seeds are common to both the IC model and the IC model with intrinsic activation – most influence models favor high degree nodes





Maximizing engagement

Ran the algorithm with three different alpha ranges and finally the scenario where the alpha values were set to be proportional to the node out-degree



Approximate Linear Model for Influence Spread under Two Modes of Activation

Approximation Model for the Activation Probabilities

$$p_A^T(i) = \alpha(i) + \left(1 - \prod_{j,(j,i)\in\mathcal{E}} \left(1 - \beta(i) w_{ij} p_A^T(j)\right)\right)$$

$$p_A^T(i) = \alpha(i) + \beta(i) \sum_{j,(j,i)\in\mathcal{E}} w_{ij} p_A^T(j)$$

$$\boldsymbol{p_A^T} = \boldsymbol{\alpha 1} + ((\boldsymbol{I} - \boldsymbol{\alpha})\,\boldsymbol{W})\,\boldsymbol{p_A^T}$$

$$\boldsymbol{p_A^T} = \boldsymbol{1}^T \boldsymbol{G}; G = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{\alpha})\boldsymbol{W}))^{-1}\,\boldsymbol{\alpha}$$

$$C_A(i) = \left(\sum_{j=1}^{N} G_{ji}\right)$$

Influence maximization vs. Equivalent Centrality

- Jaccard Index : Ratio of the number of elements in the intersection of the two sets and the number of elements in the union of the two sets

- Rank Based Overlap (RBO) – gives importance to not only the overlaps and rank alignment, but also by appropriately weighting the top ranks more than the lower ranks (higher the score, better the correlation)

Correlation type	Input	$k = 10$	$k = 20$	$k = 30$	$k = 50$
Jaccard	PBlogs	0.538	0.818	0.875	0.818
RBO	PBlogs	0.817	0.846	0.851	0.868
Jaccard	LFR1000	0.818	0.905	0.765	0.818
RBO	LFR1000	0.979	0.963	0.947	0.937

Cyber Risk Modeling

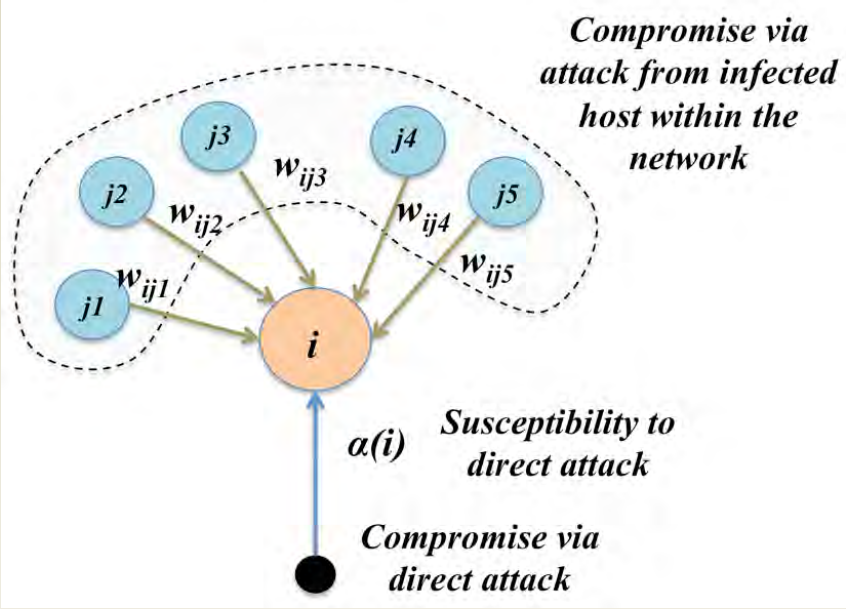
Analyst risk assessment scenario

Quickly assess the changing landscape of enterprise cyber risk encompassing multiple mechanisms

Needs to consider	Enterprise graph
	Machine vulnerabilities
	Behavior of associated human users
	Possibly include confidence bounds on the metrics

"What is the risk potential of an enterprise laptop used by an employee who regularly clicks web-links in personal emails and downloads trial software from the Internet, on three specific high-asset enterprise workstations in particular and the enterprise as a whole?"

Modes of Risk Proliferation



Direct : The host can be infected with malware via an email attachment or by visiting a malicious website or via an infected USB stick.

Network : A host that is in the neighborhood of the host under scrutiny is first compromised and then the host under consideration is compromised through existing vulnerabilities

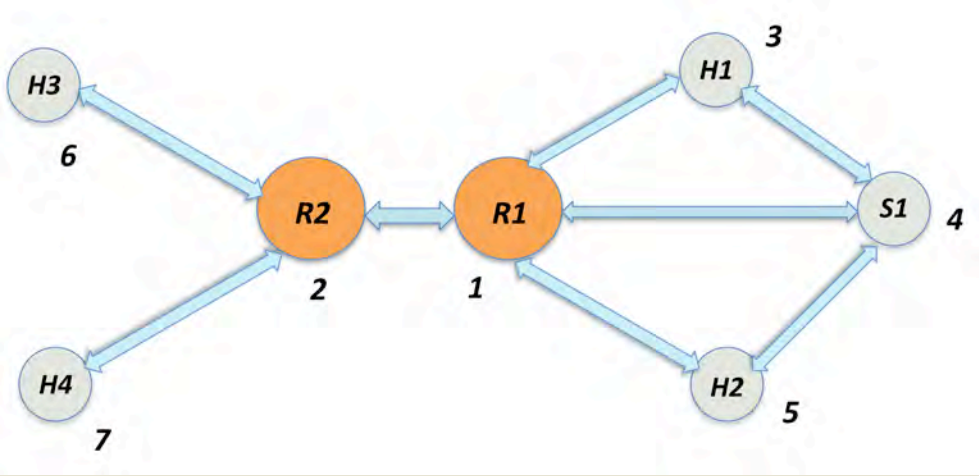
Arun V. Sathanur and David J. Haglin, "A Novel Centrality Measure for Network-wide Cyber Vulnerability Assessment," *IEEE Conference on Technologies for Homeland Security*, Waltham, MA, May 2016

Model parameters

- Node model parameters $\alpha(i)$
  - Represents the susceptibility to direct attack
    - Risky user behaviors
    - Host exposure to the Internet
    - Host position on the cyber graph

- Edge model parameters w_{ij}
  - The probability that host i, is compromised by an already compromised host j where (i,j) is an edge.
    - Based on the existence of vulnerabilities that are lined up through network services (example *open ftp ports* and *ftp services running* and *vulnerability in the ftp server*)
    - This parameter essentially summarizes the local attack graph between host i and host j.

Toy Example



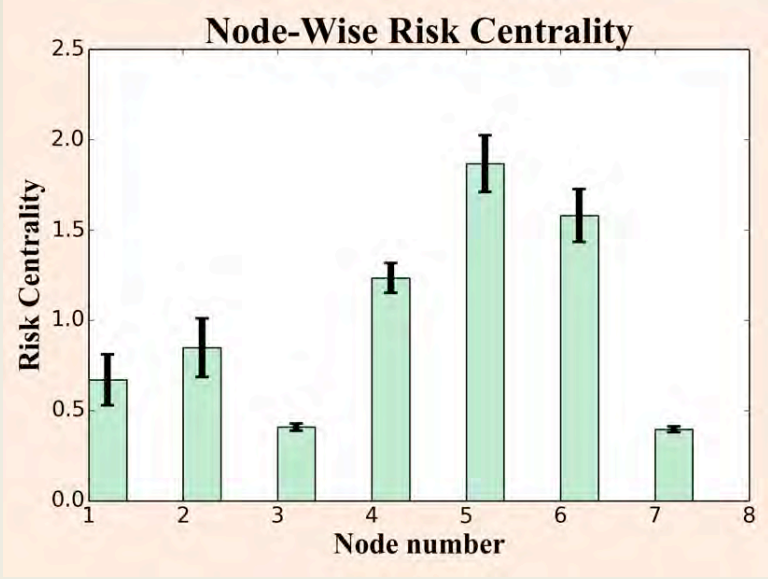
Two Routers (R); One Server (S) and Four workstations / laptops (H1-H4)

Based on their roles on the network and the similarities in the vulnerabilities between the machines, values are assigned to *α* and *W entries.*

Nodes with a high susceptibility to direct attack/compromise and that are central or connect to central nodes tend to have higher risk centrality.

These observations are compatible with the first order approximation for the risk centrality.

Topologically central nodes R1 and R2 figure low on the list because of low susceptibilities to direct attack/compromise (low α values).



Conclusions and Future Work

Discussion

- Influence dynamics on a number of complex networks can be better modeled by considering intrinsic and influenced activation modes
- Demonstrated the role of the intrinsic activation parameter α in shaping the dynamics on a social network
- Derived a centrality metric based on a linear approximation model for fast analyses and optimization
- Motivated the approach for application in a cyber risk assessment use-case

Future work

- Generalization of the intrinsic-influenced activation mechanisms to other diffusion models
- Scaling of the algorithms to 10s of millions nodes
  - Accelerated Submodular Optimization
  - Accelerated Reachability Computations
- Applying the methodologies to real-world cyber datasets