Anomalous Node Detection in Homophilic Networks with Communities of Varying Size

Juan Camilo Campos

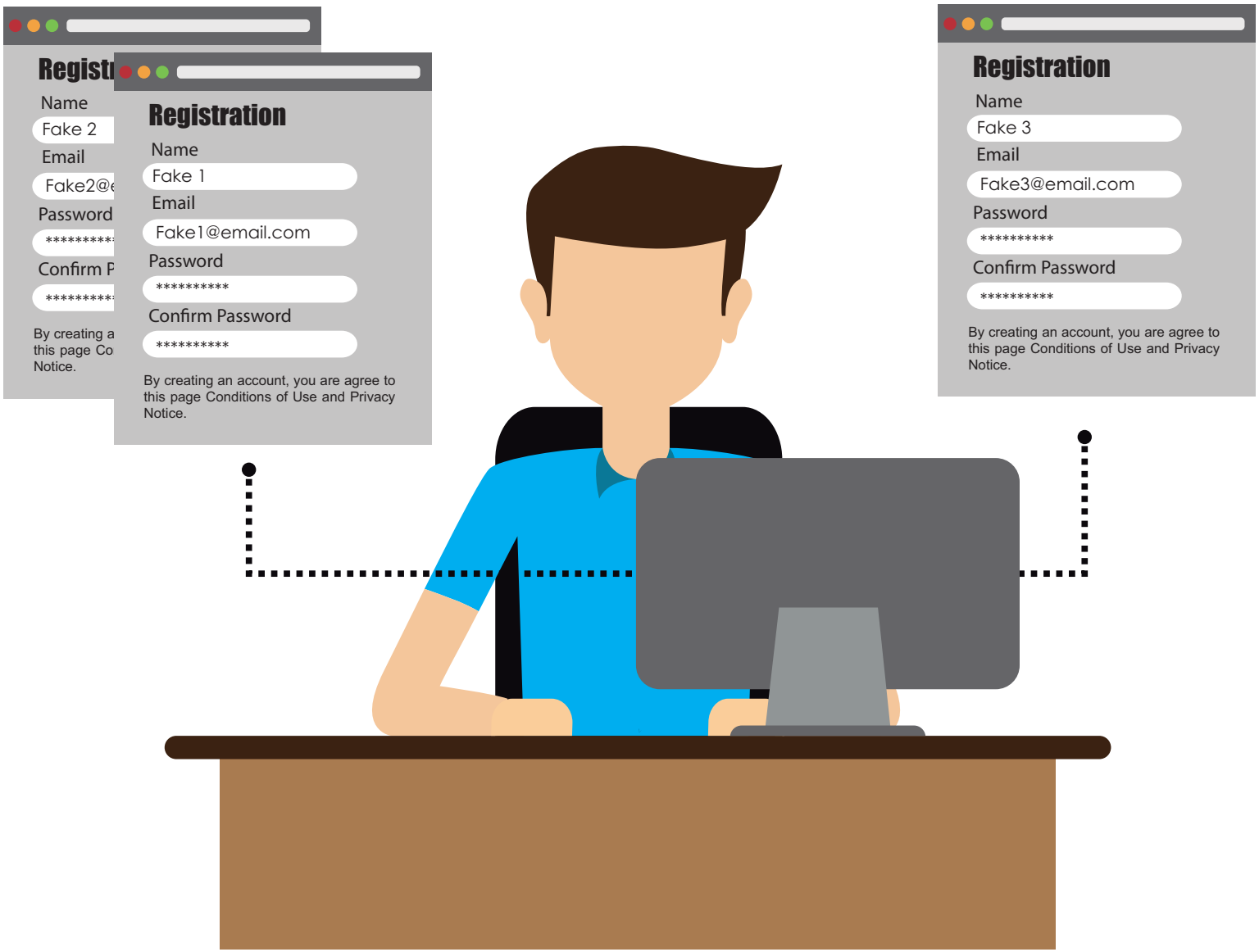
Dpt. Electrical Engineering and Computer Science
Pontificia Universidad Javeriana
Santiago de Cali, Colombia

May 2017

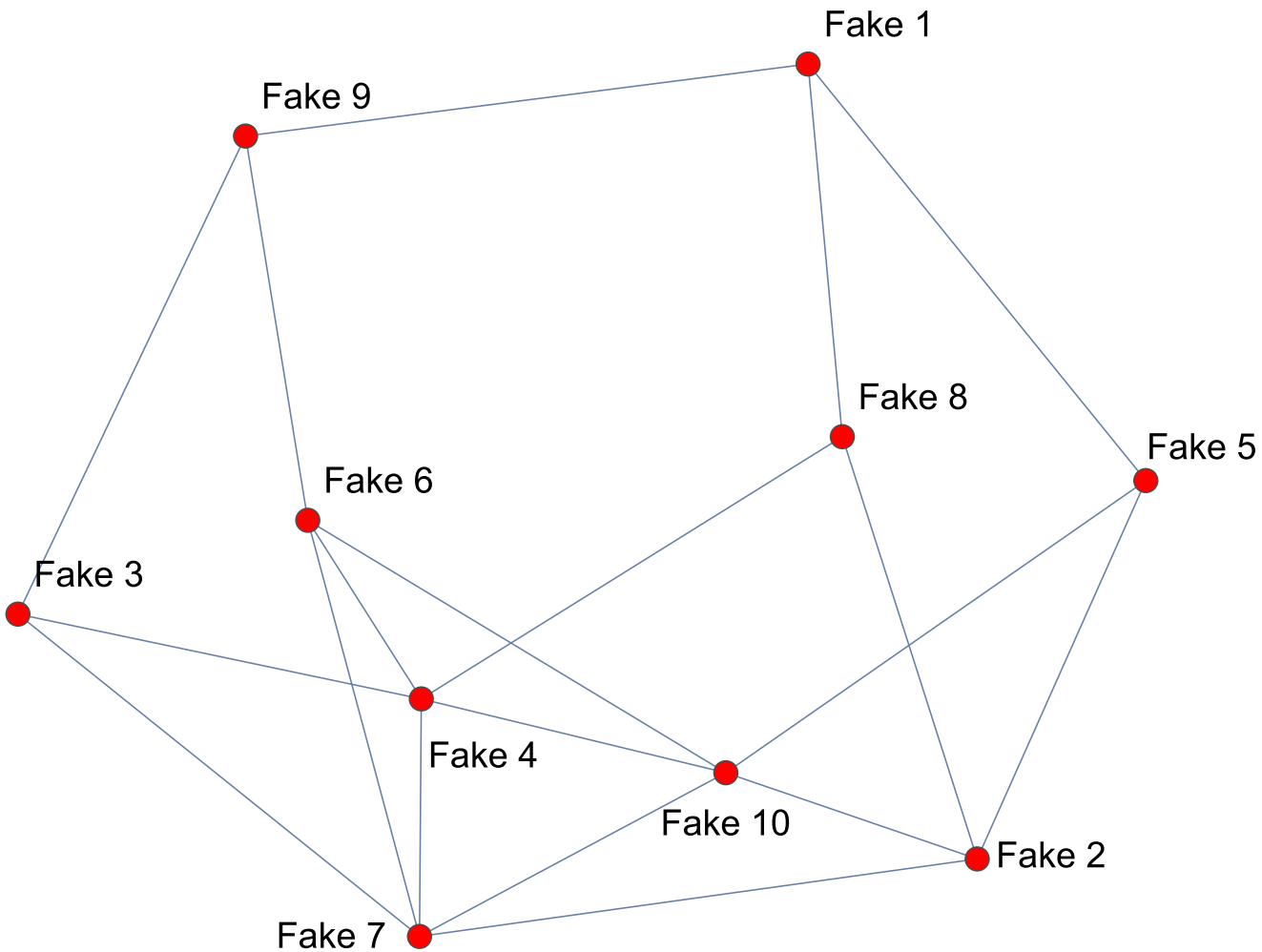
Motivation

- Large interaction platforms
- Hundred of thousands of transactions
- Size $\rightarrow$ easy target for fraudsters (anomalous nodes)
- Anomalous node: someone trying to deceive regular user behavior

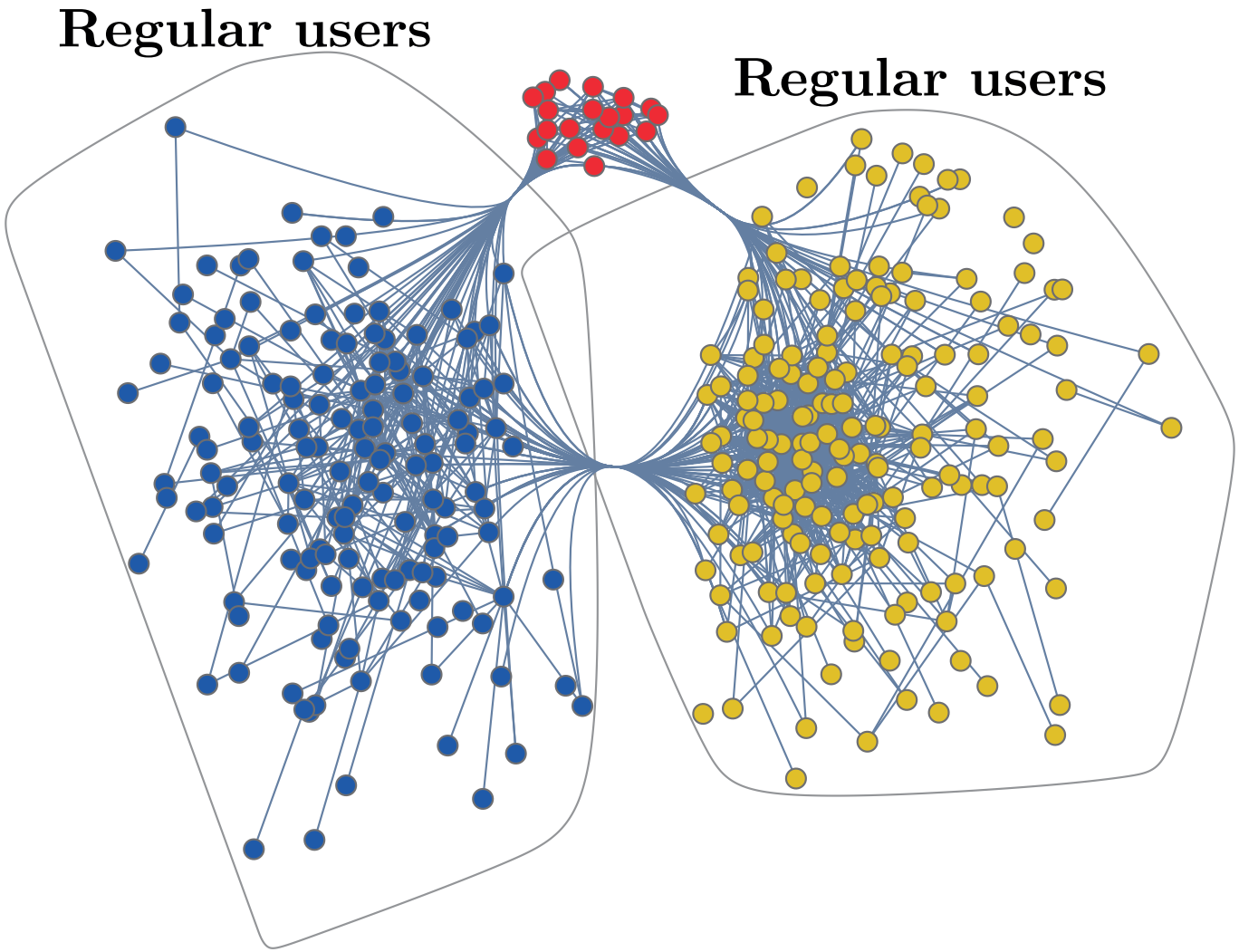
Scenario


Registration
Name
Fake 1
Email
Fake1@email.com
Password
****
Confirm Password
****
By creating an account, you are agree to this page Conditions of Use and Privacy Notice.
Registration
Name
Fake 3
Email
Fake3@email.com
Password
****
Confirm Password
****
By creating an account, you are agree to this page Conditions of Use and Privacy Notice.

Scenario


Fake 1
Fake 9
Fake 8
Fake 5
Fake 6
Fake 3
Fake 4
Fake 10
Fake 2
Fake 7

Scenario



Overview

Previous concepts

Homophily

Previous concepts

Random Link Attacks (RLAs)

Problem Definition

Given:

- A network that is divided into two communities (with strong homophilic relationships); and
- Some anomalous nodes who perform RLAs.

We want to:

- Characterize the expected cohesion indices for varying community sizes, and
- Find anomalous node, i.e., users who are performing RLAs.

Related work

- K. Guerrero and J.Finke, "On the formation of community structures from homophilic relationships", *IEEE Proceedings of the American Control Conference*, (Montreal, Canada), pp. 5318-5323, June 2012

- X. Ying, X. Wu, and D. Barbará, "Spectrum based fraud detection in social networks," *Proceedings of the IEEE International Conference on Data Engineering*, (Hannover, Germany), pp.912-923, April 2011

Characterize dynamics

The model

$G(t) = (N, A(t))$: network at time t

$N = \{1, ..., n\}$: set of nodes

$A(t), \{i, j\} \in A(t)$ if node i links to node j: set of edges

$M(t)$, $m_{i,j}(t) \in \{0, 1\}$: adjacency matrix

N_1, N_2: sets of regular nodes

N_0: set of anomalous nodes

$g_i : N \to \{0, 1, 2\}$: type of a node

The model

$A_i(t) = \{\{j', j\} \in A(t) : j' = i\}$: neighborhood

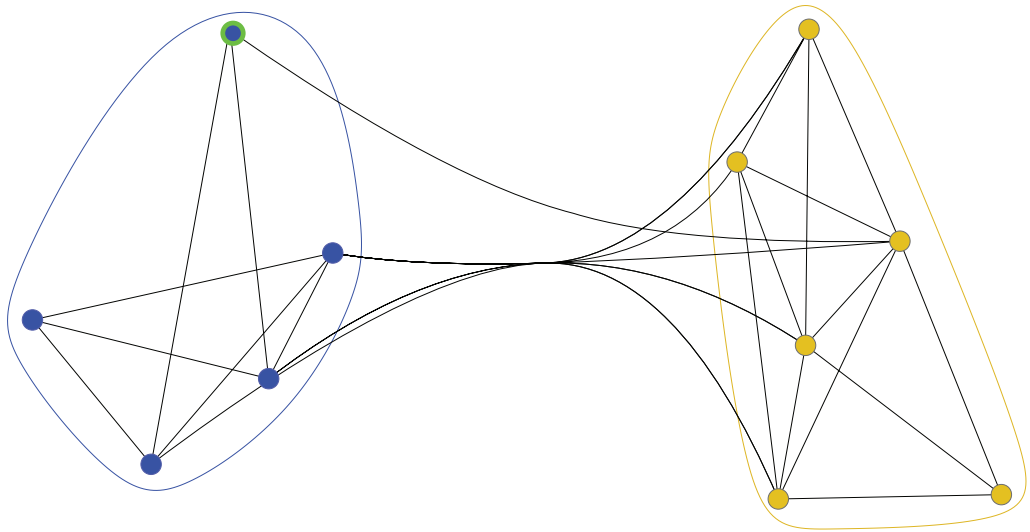
$R_i(t) \subseteq A_i(t)$: subset of edges established by a node

$r = |R_i(t)|$: edges that each node establishes

$k_i^{\delta}(t)$: number of links to nodes of the same type

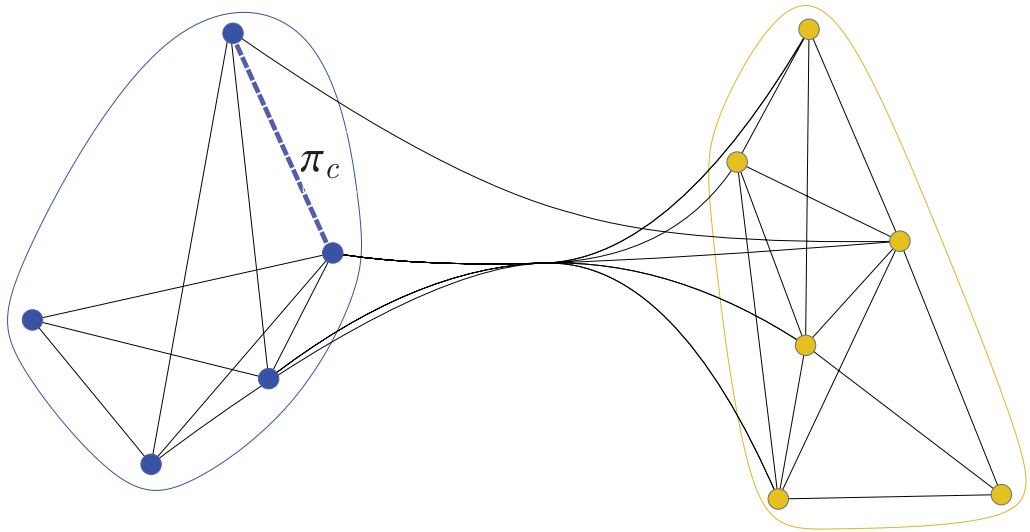
$k_i(t) = |A_i(t)|$: degree of a node

regular node



$$w_c = \begin{cases} w & \text{if } g_i = g_c, \\ 1 - w & \text{if } g_i \neq g_c. \end{cases}$$

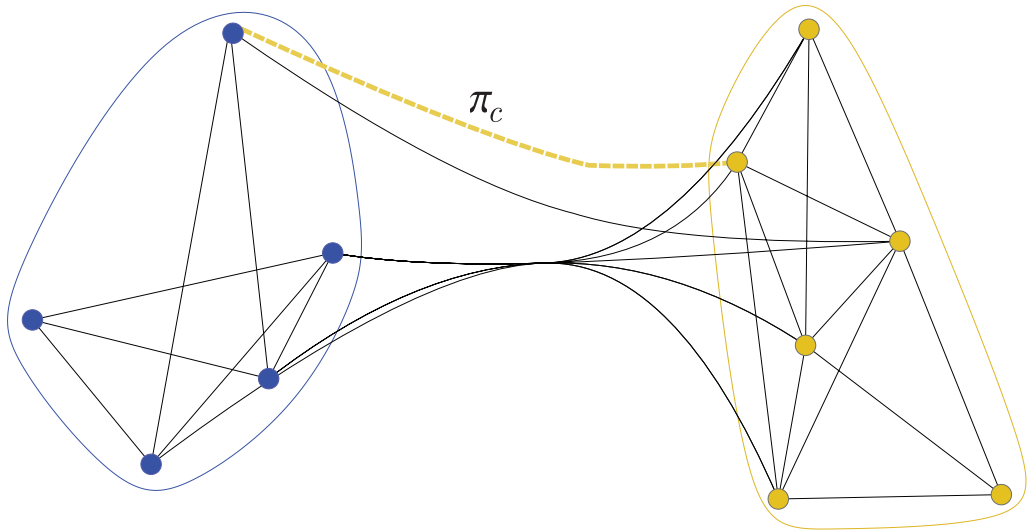
regular node



$$\pi_c(t) = w_c\, k_c(t)\, \frac{1}{\sum\limits_{\{i,j\}\in A_i^c(t)} w_j k_j(t)}$$

$$w_c = \begin{cases} w & \text{if } g_i = g_c, \\ 1-w & \text{if } g_i \neq g_c. \end{cases}$$

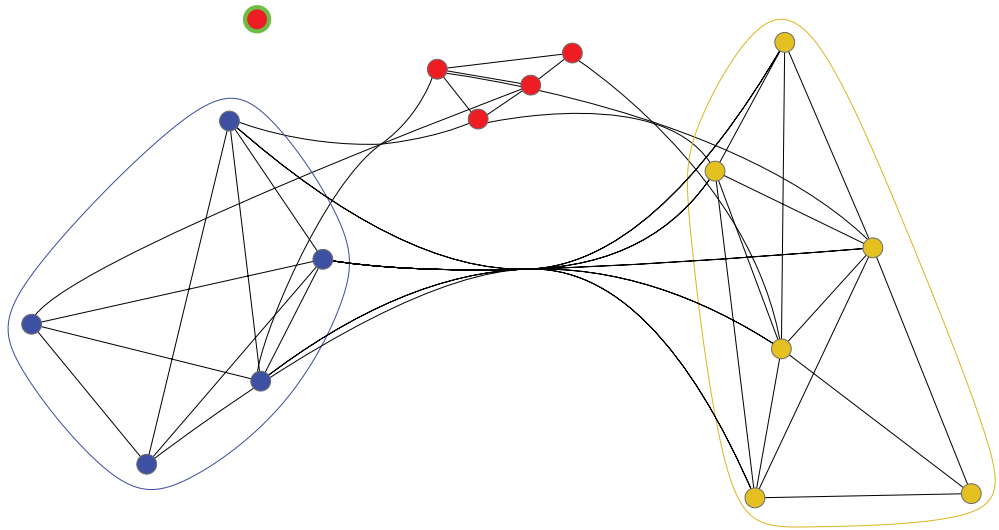
regular node



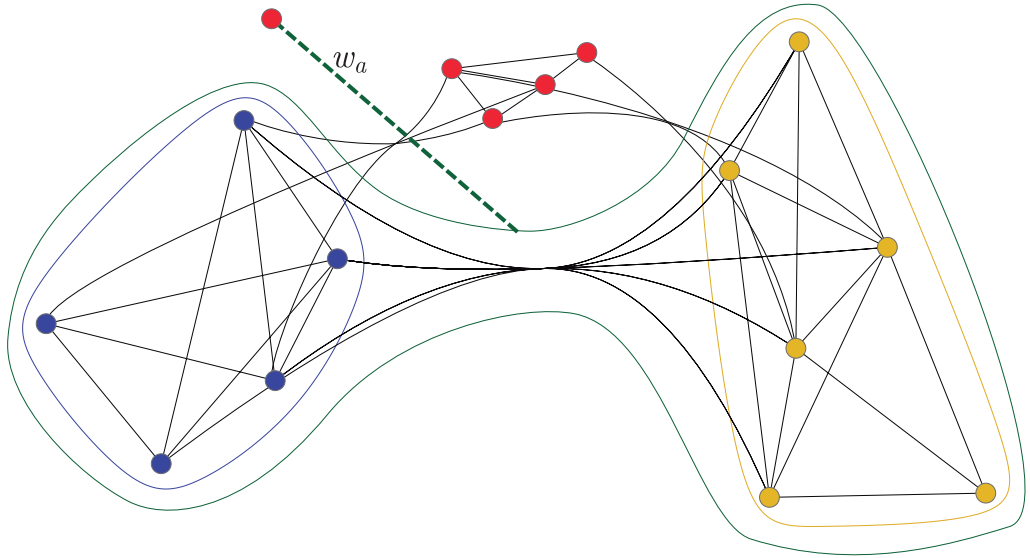
$$\pi_c(t) = w_c\, k_c(t)\, \frac{1}{\sum\limits_{\{i,j\}\in A_i^c(t)} w_j k_j(t)}$$

$$w_c = \begin{cases} w & \text{if } g_i = g_c, \\ 1-w & \text{if } g_i \neq g_c. \end{cases}$$

anomalous node

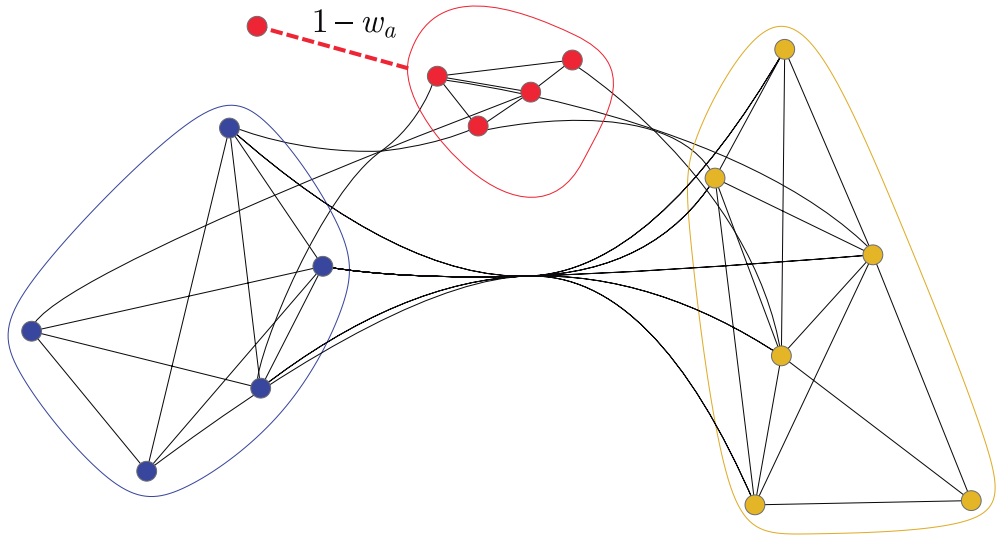


anomalous node



$$\pi_c = \begin{cases} \dfrac{1 - w_a}{n_0} & \text{if } g_c = 0, \\ \dfrac{w_a}{n_1 + n_2} & \text{if } g_c \in \{1, 2\}. \end{cases}$$

anomalous node

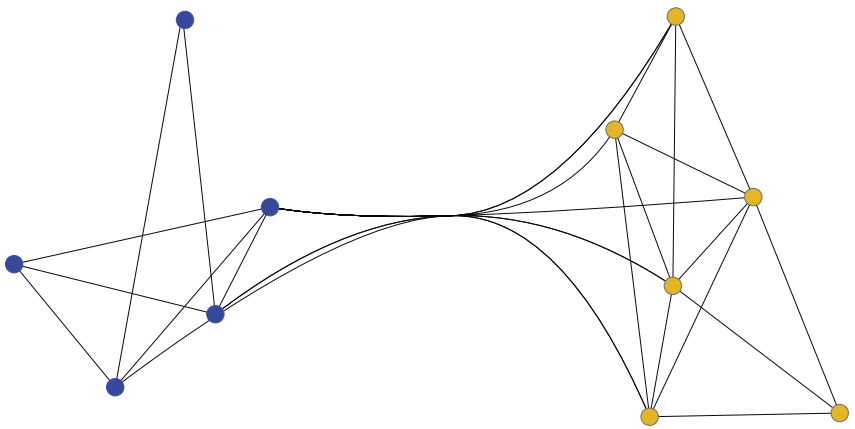


$$\pi_c = \begin{cases} \dfrac{1 - w_a}{n_0} & \text{if } g_c = 0, \\ \dfrac{w_a}{n_1 + n_2} & \text{if } g_c \in \{1, 2\}. \end{cases}$$

Topological Measures

Cohesion index

$$h_\delta(t) = \frac{1}{n_\delta} \sum_{i \in N_\delta} \frac{k_i^\delta(t)}{k_i(t)}$$

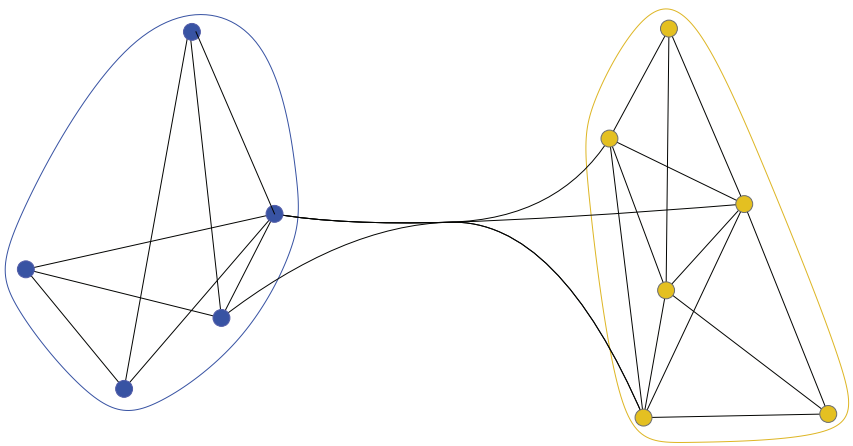


The average proportion of neighbors of the same type

Topological Measures

Community modularity

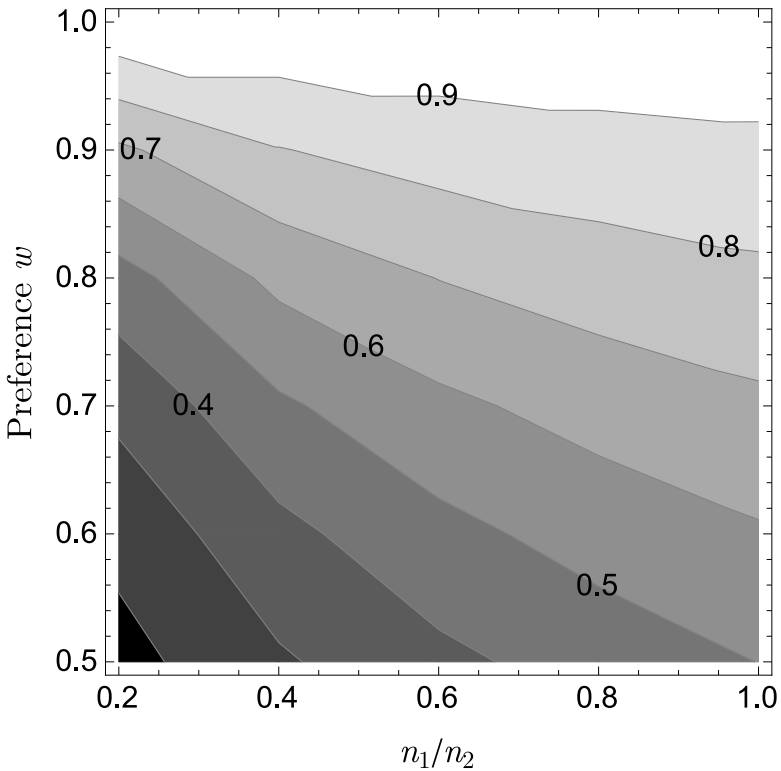
$$q(t) = \sum_{\delta=1}^{2} \left(\frac{|\{i,j\} \in A(t) : g_i = g_j = \delta|}{|A(t)|} - \frac{|\{i,j\} \in A(t) : g_i = \delta \text{ or } g_j = \delta|^2}{|A(t)|^2} \right)$$



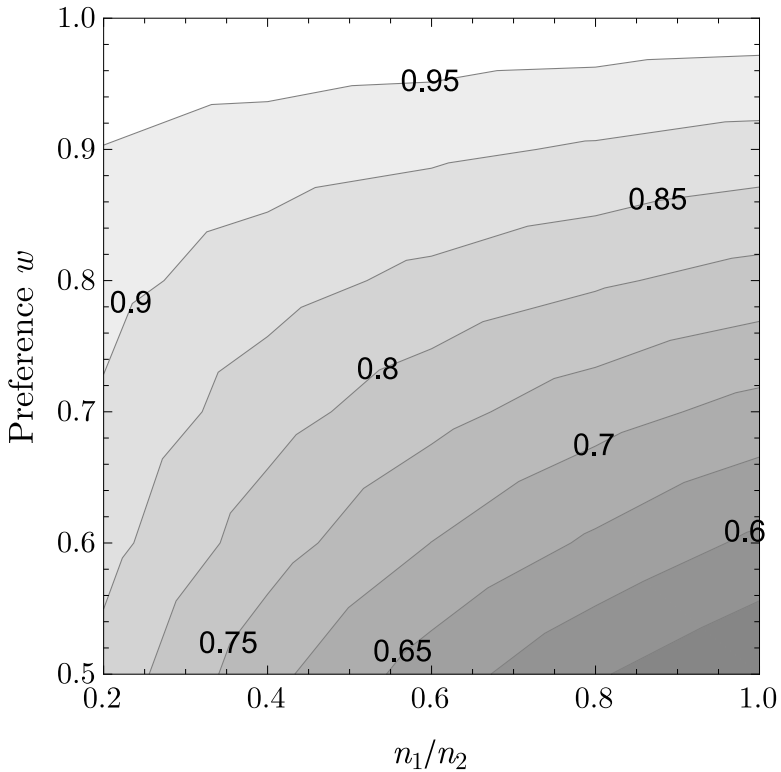
Modularity is based on the number of edges within communities compared to the number of edges between them

Topological properties

Expected cohesion index



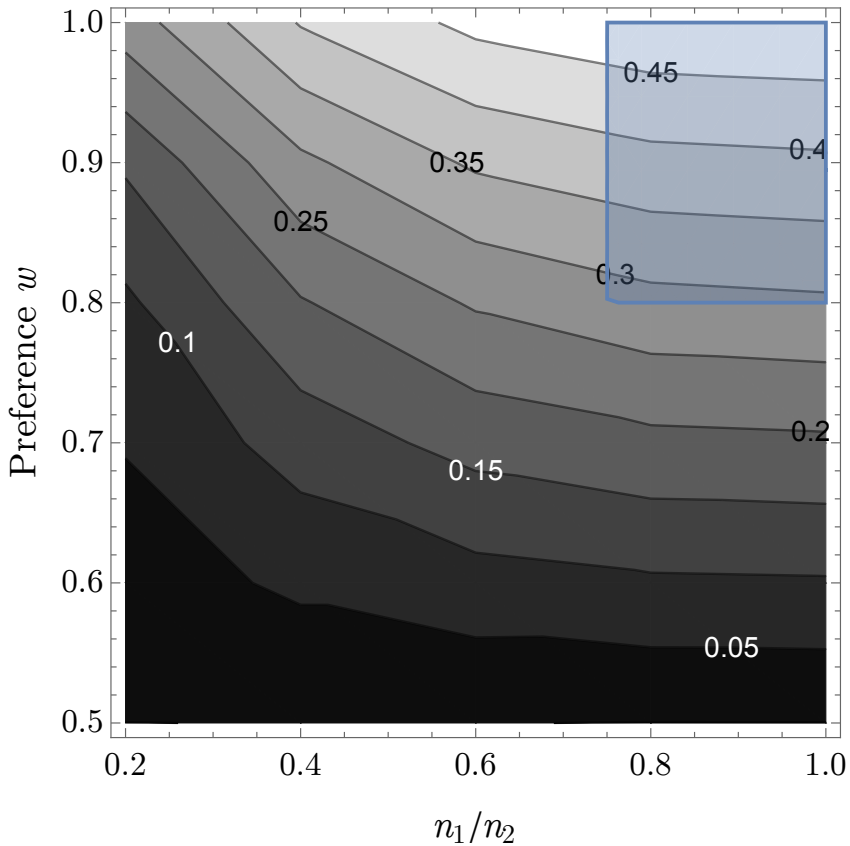
minority group



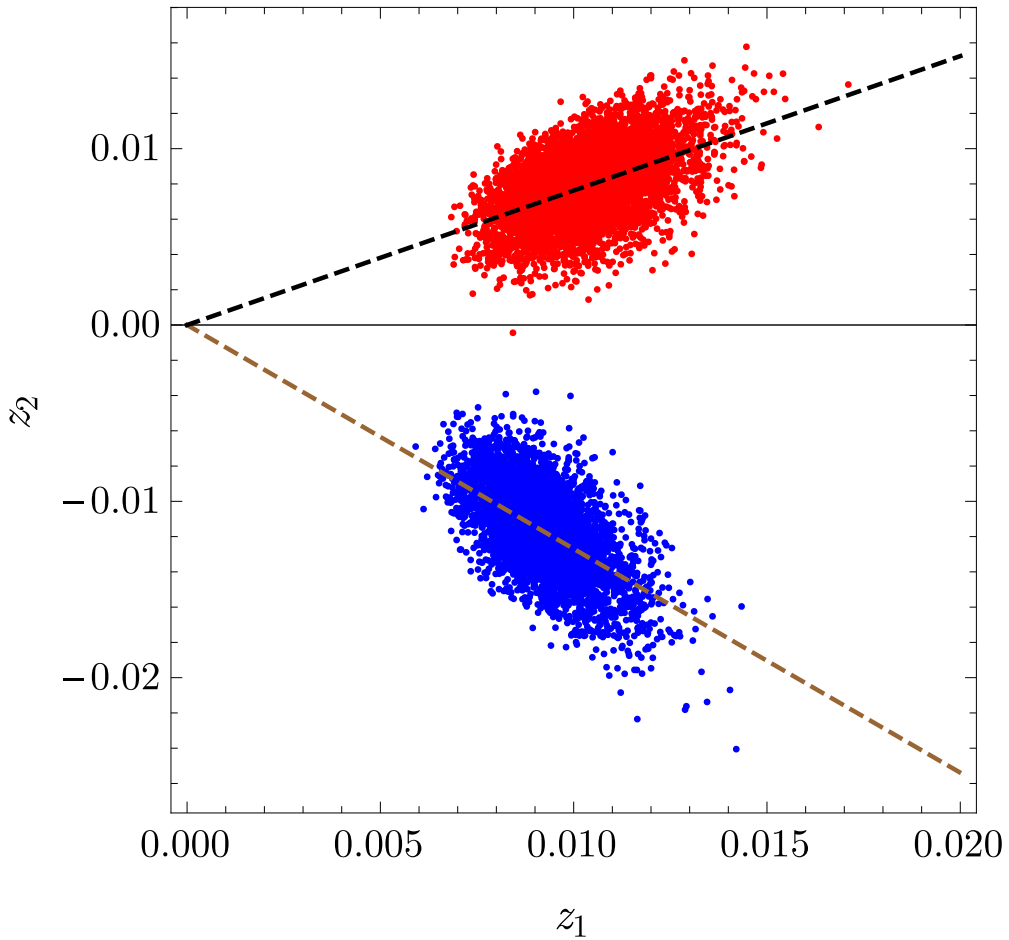
majority group

Topological properties

Average community modularity



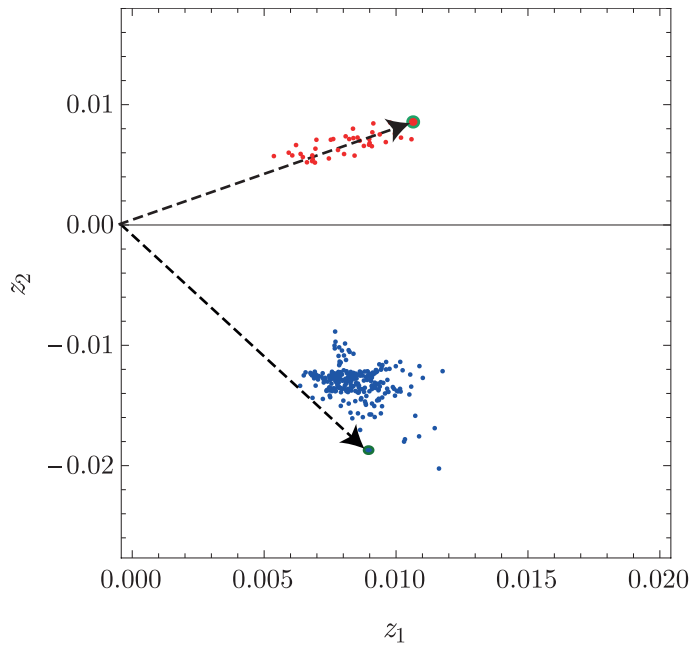
Spectral properties



Detection (Approach B)

Edge-non-randomness

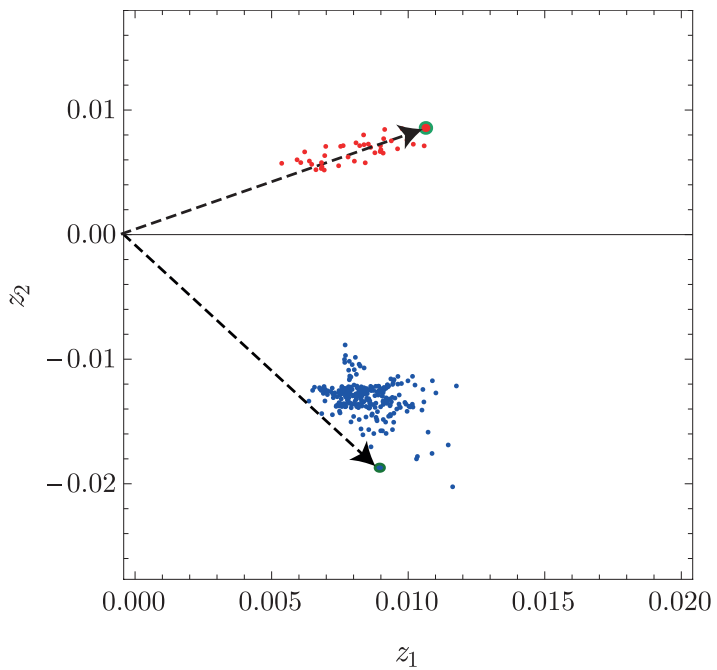
$$f(i,j) = ||\alpha_i||_2 ||\alpha_j||_2 \cos(\alpha_i, \alpha_j)$$



Detection (Approach B)

Edge-non-randomness

$$f(i,j) = ||\alpha_i||_2 ||\alpha_j||_2 \cos(\alpha_i, \alpha_j)$$

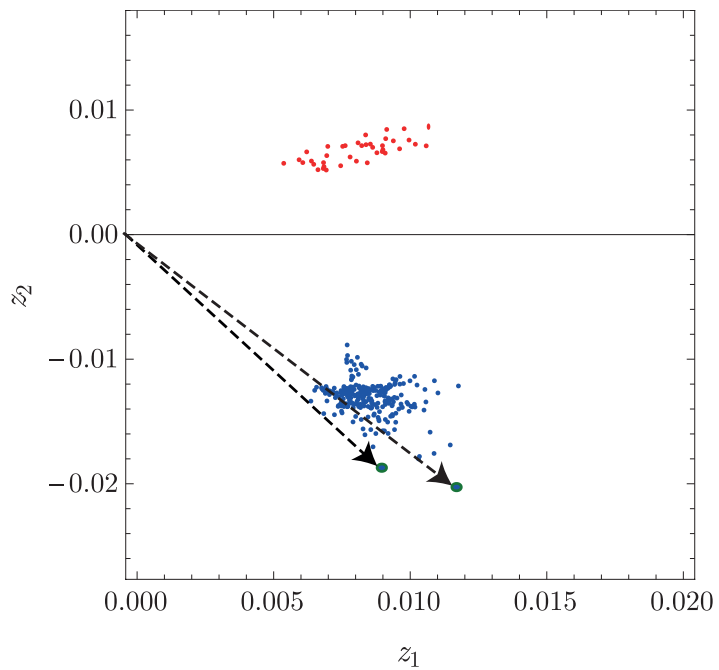


$$\cos(\alpha_i, \alpha_j) \approx 0$$

Detection (Approach B)

Edge-non-randomness

$$f(i,j) = ||\alpha_i||_2 ||\alpha_j||_2 \cos(\alpha_i, \alpha_j)$$



$$\cos(\alpha_i, \alpha_j) \approx 1$$

Detection (Approach A)

Identifying Suspects

- Degree of membership to well-defined communities based on node-non-randomness

$$f_i(t) = \sum_{j \in A_i(t)} f(i,j) = \sum_{j=1}^{2} \lambda_j(t)\ z_{ji}^2(t)$$

- Suspect if

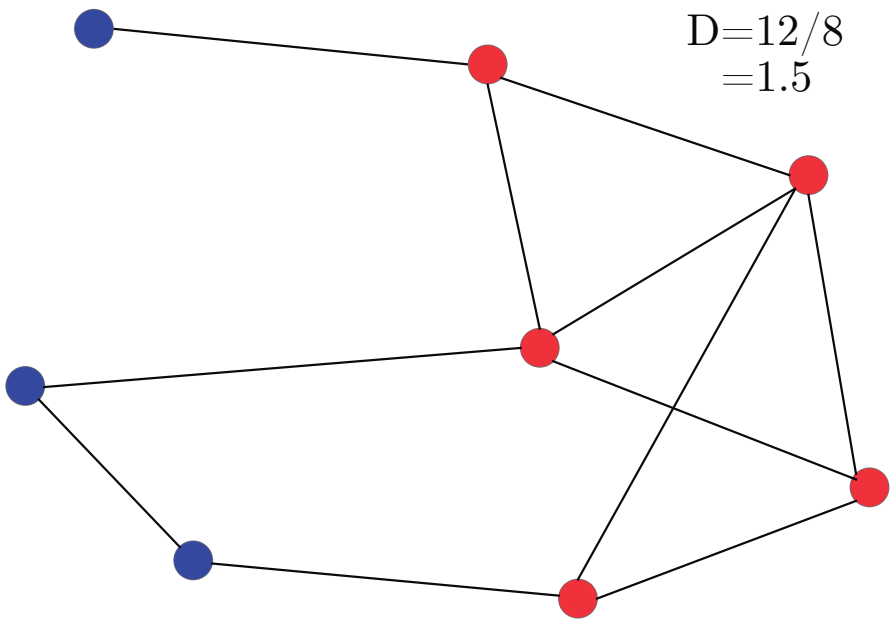
$$f_i \ \leq \ B_i^E + \beta (B_i^V)^{1/2}$$

B_i^E and B_i^V: upper bounds of the expected value and variance

Detection (Approach A)

Detecting anomalous nodes

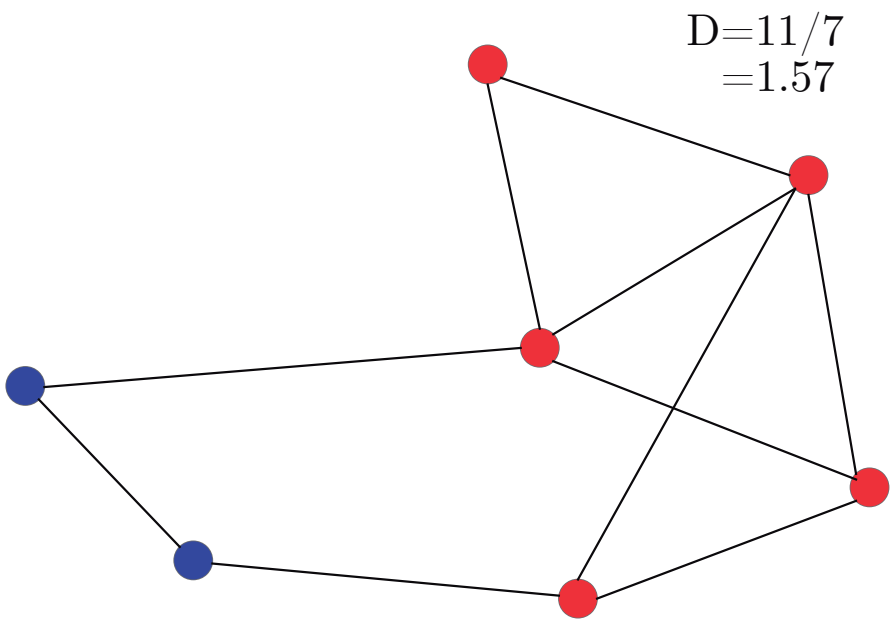
Most-dense subgraph (number of edges/number of nodes)



Detection (Approach A)

Detecting anomalous nodes

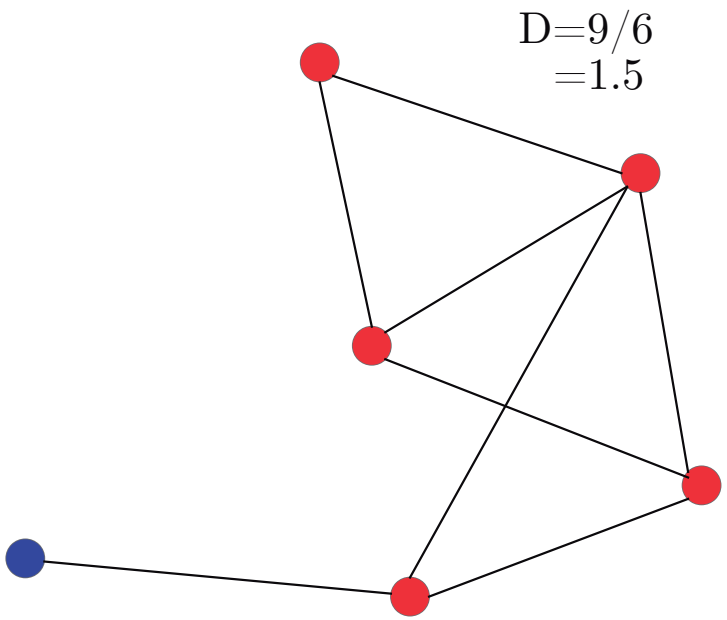
Most-dense subgraph (number of edges/number of nodes)



Detection (Approach A)

Detecting anomalous nodes

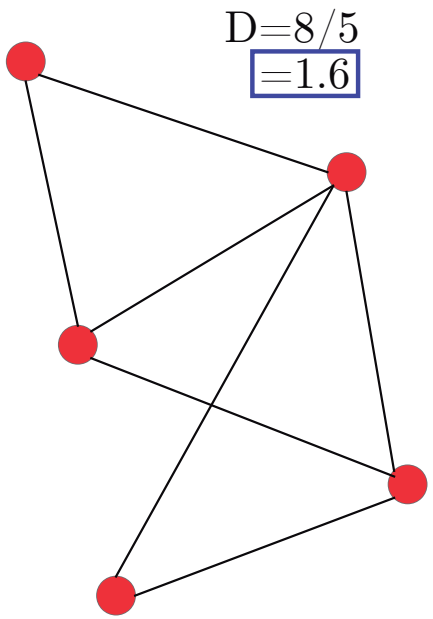
Most-dense subgraph (number of edges/number of nodes)



Detection (Approach A)

Detecting anomalous nodes

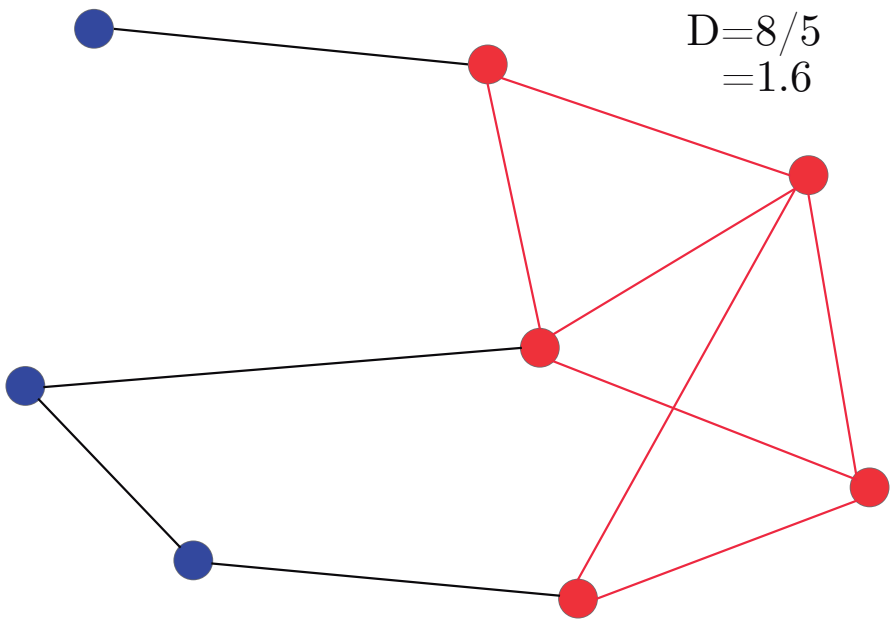
Most-dense subgraph (number of edges/number of nodes)



Detection (Approach A)

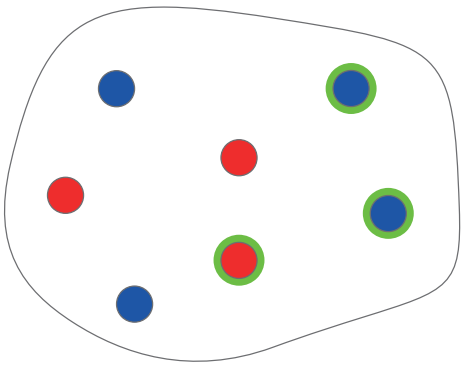
Detecting anomalous nodes

Most-dense subgraph (number of edges/number of nodes)



Algorithm performance

Performance Measures



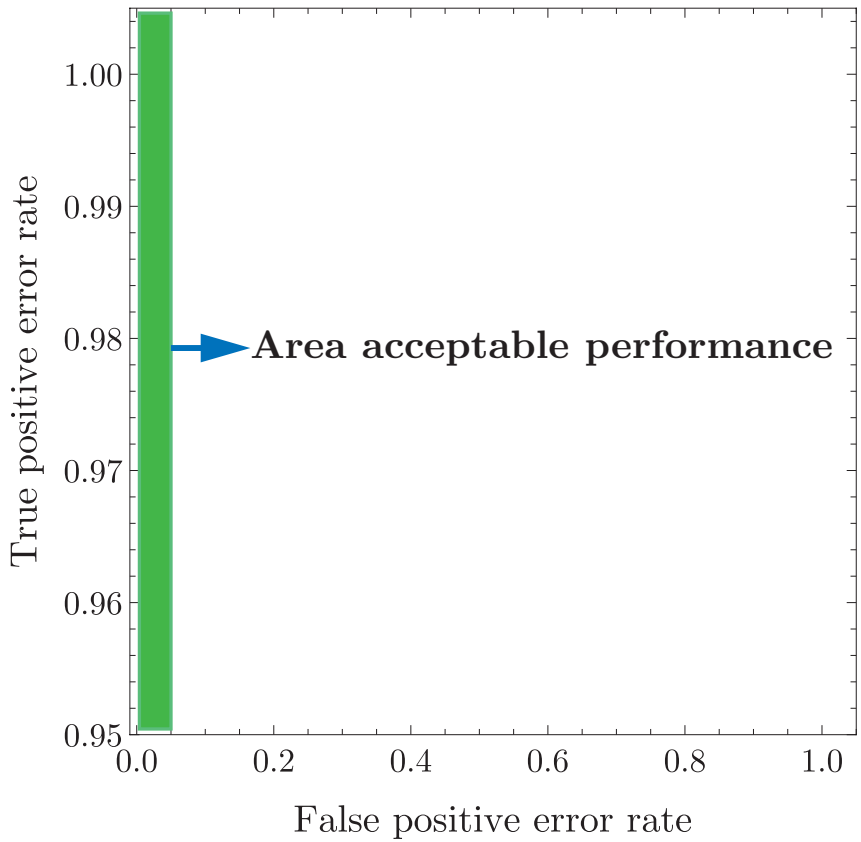
$e_1 = 2/3$
$e_2 = 1/3$

- False positive error rate (e_1): number of regular nodes accused as anomalous nodes over the total number of accused nodes

- True positive error rate (e_2): number of anomalous nodes detected over the total number of anomalous nodes

Algorithm performance

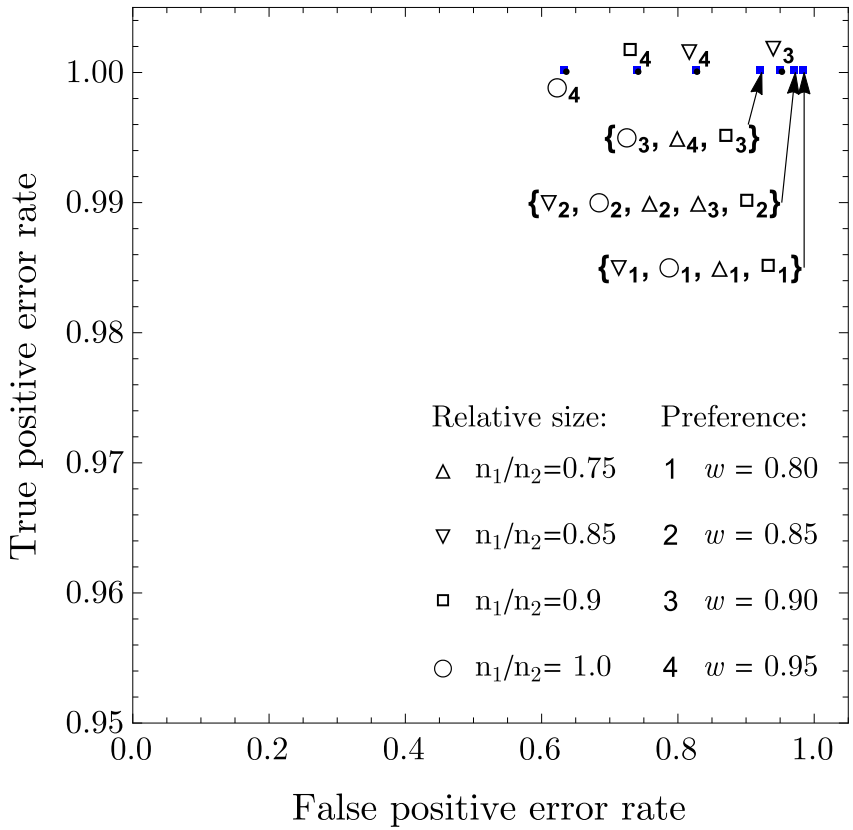
Performance Measures



$$e_1 \leq 0.05 \text{ and } e_2 \geq 0.95$$

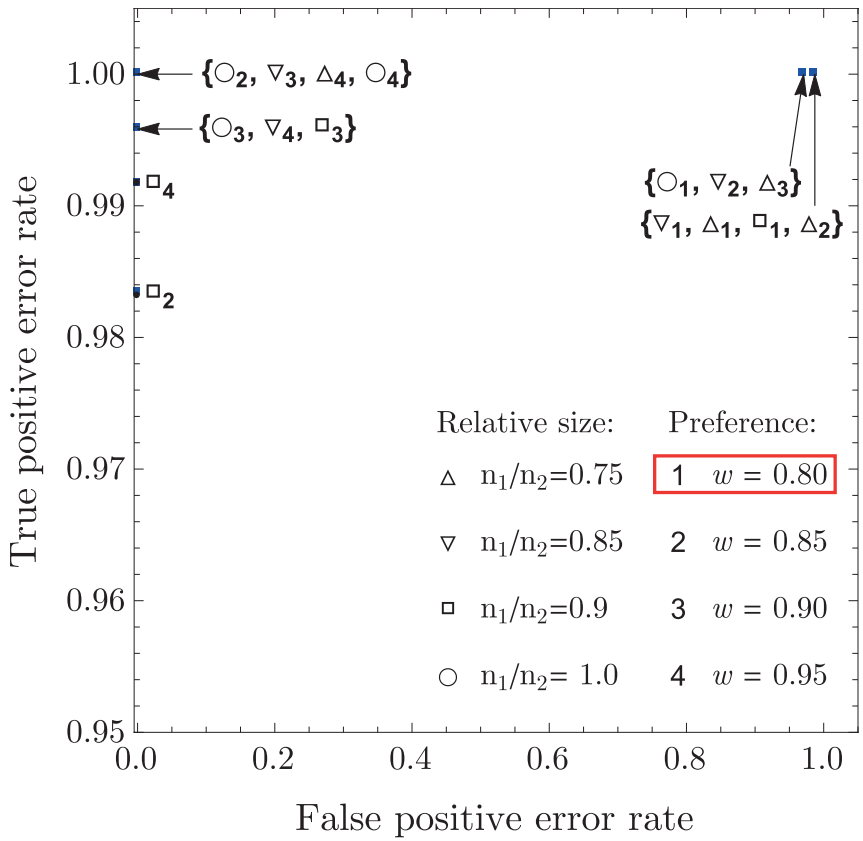
Algorithm performance (Approach A)

Identification of suspects



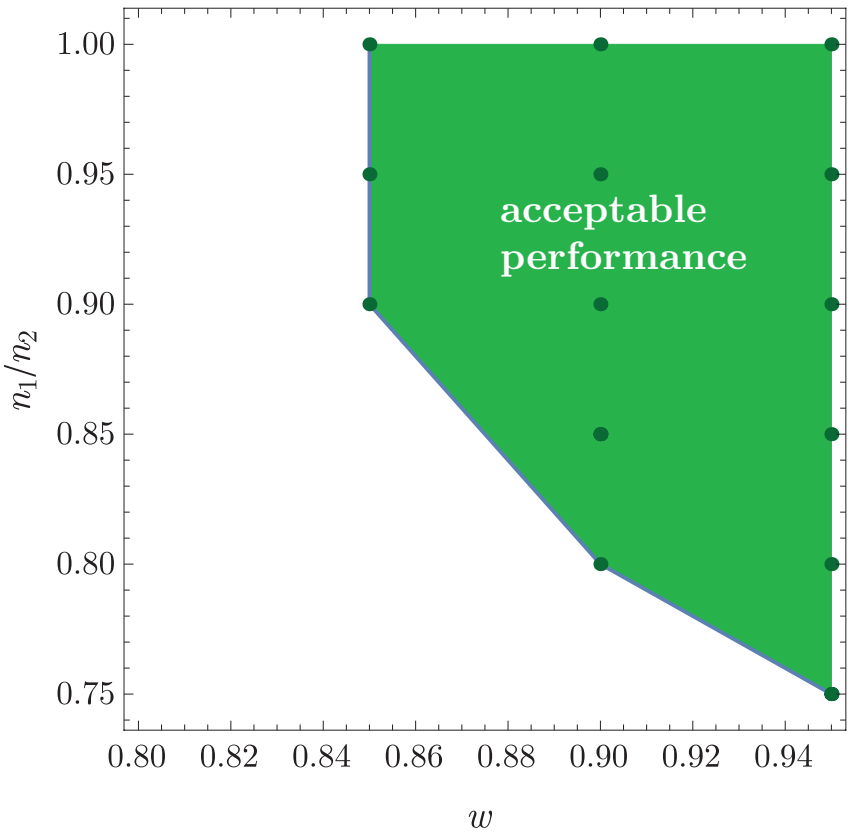
Algorithm performance

Detection of anomalous nodes



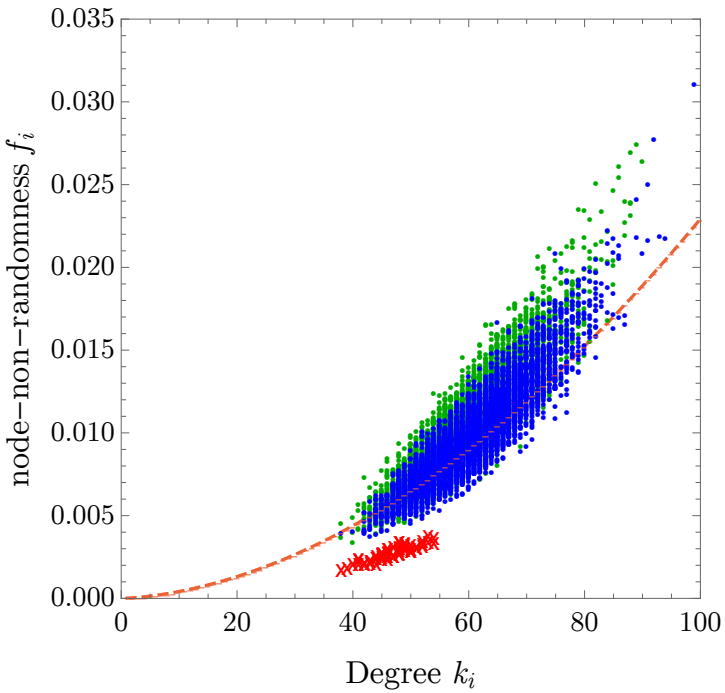
Algorithm performance

Area of acceptable performance

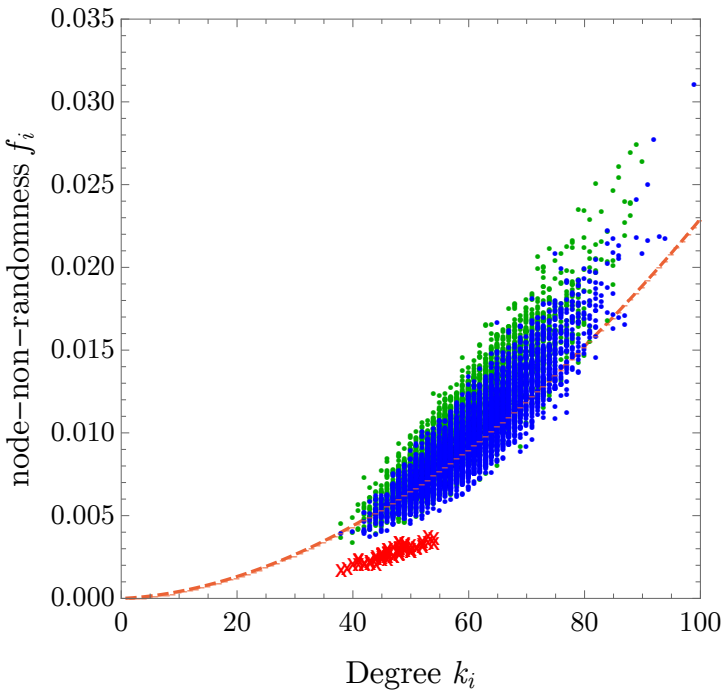


Performance of the Approach A for all generated networks

Node-non-randomness



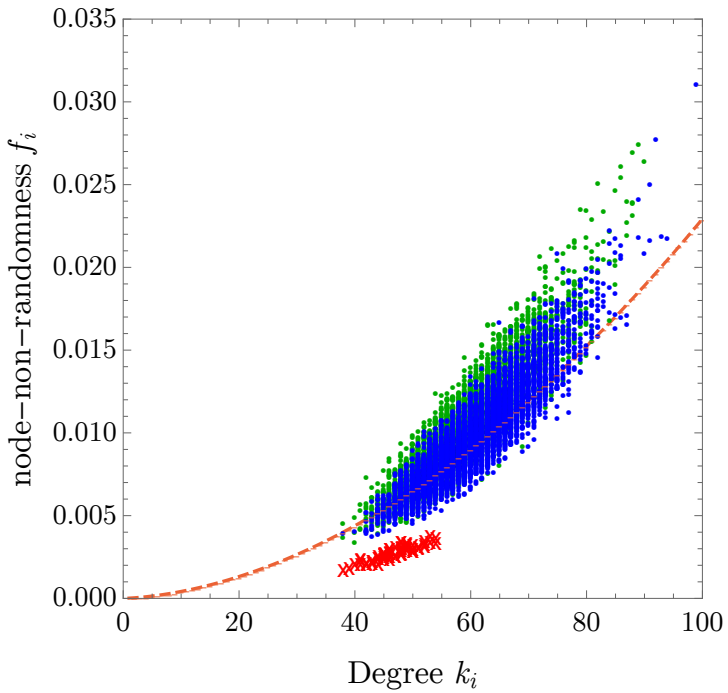
Node-non-randomness



Suspects distinguishable from regular nodes

design parameter
$\beta = 2$

Node-non-randomness



Suspects distinguishable from regular nodes

design parameter
$\beta \neq 2$

β is sensitive to the group size, and to the preference level

Detection (Approach B)

Identifying Suspects

Expected spectral coordinates + variance

$$E[z_{ji}(t)] \leq \frac{k_i(t)\, E[z_j(t)]}{\lambda_j(t)} = U_{ji}^{E}(t)$$

$$V[z_{ji}(t)] \leq \frac{k_i(t)}{n}\left(1 - \frac{k_i(t)}{n}\right)\frac{1}{\lambda_j(t)^2} = U_{ji}^{V}(t)$$

Detection (Approach B)

Identifying Suspects

$$z_{ji}(t) \in \left[U^E_{ji}(t) - \epsilon\, U^V_{ji}(t), U^E_{ji}(t) + \epsilon\, U^V_{ji}(t)\right]$$

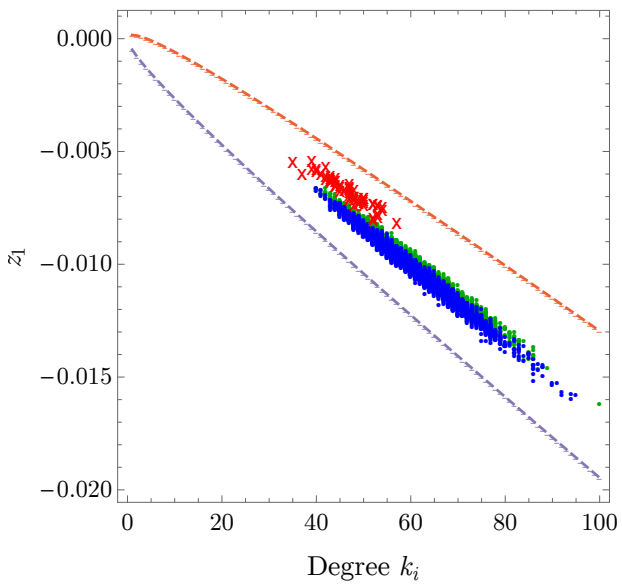
design parameter $\epsilon = 2$

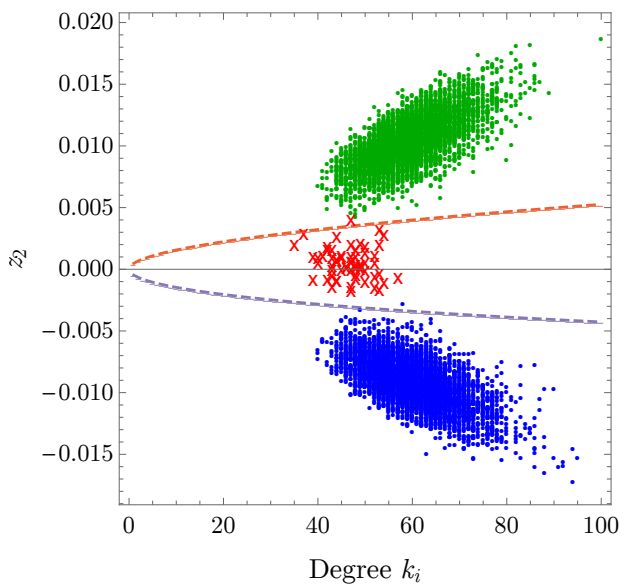
Detection (Approach B)

Identifying Suspects

$$z_{ji}(t) \in \left[U^E_{ji}(t) - \epsilon\, U^V_{ji}(t), U^E_{ji}(t) + \epsilon\, U^V_{ji}(t)\right]$$

design parameter $\epsilon = 2$



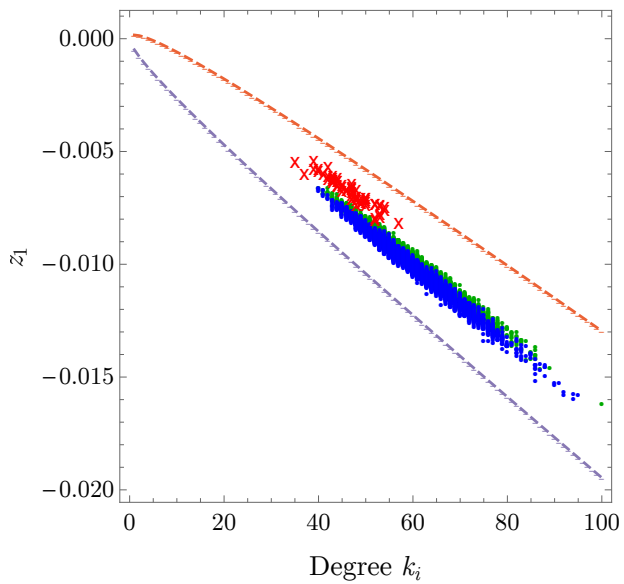


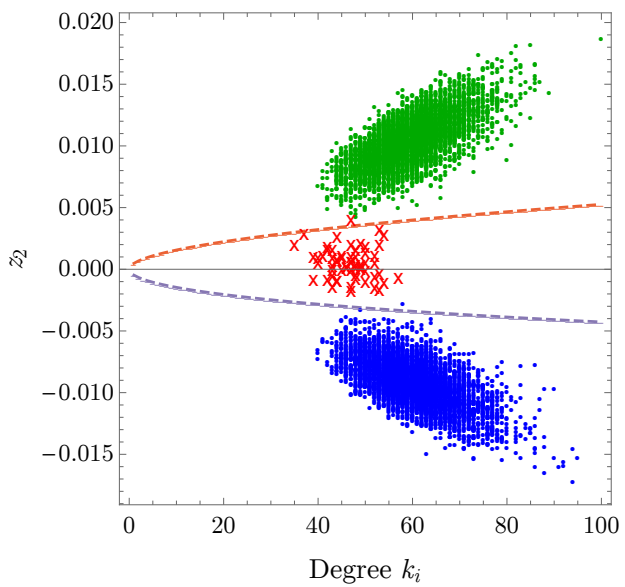
Detection (Approach B)

Identifying Suspects

$$z_{ji}(t) \in \left[U^E_{ji}(t) - \epsilon\, U^V_{ji}(t), U^E_{ji}(t) + \epsilon\, U^V_{ji}(t)\right]$$

design parameter $\epsilon = 2$

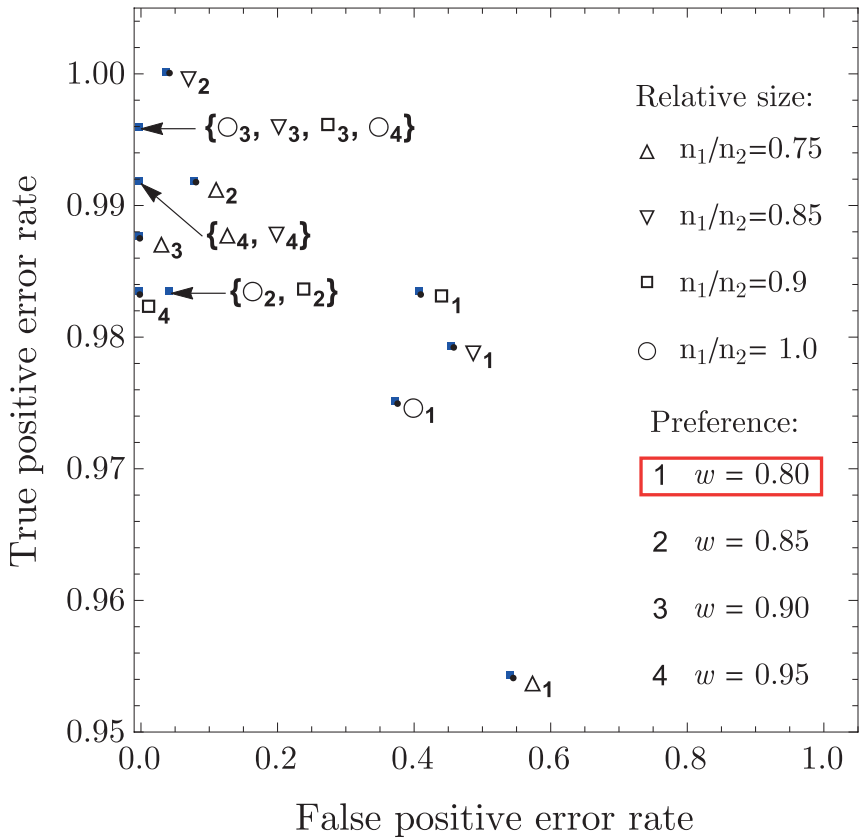




Detecting anomalous nodes (as for Approach A)

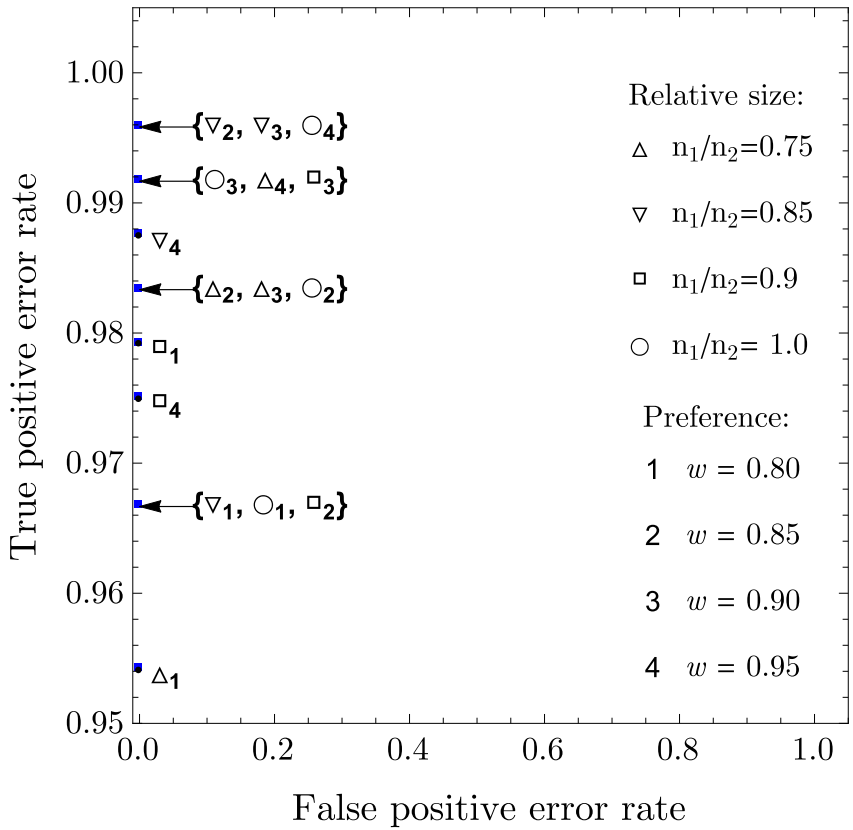
Algorithm performance (Approach B)

Identification of suspects



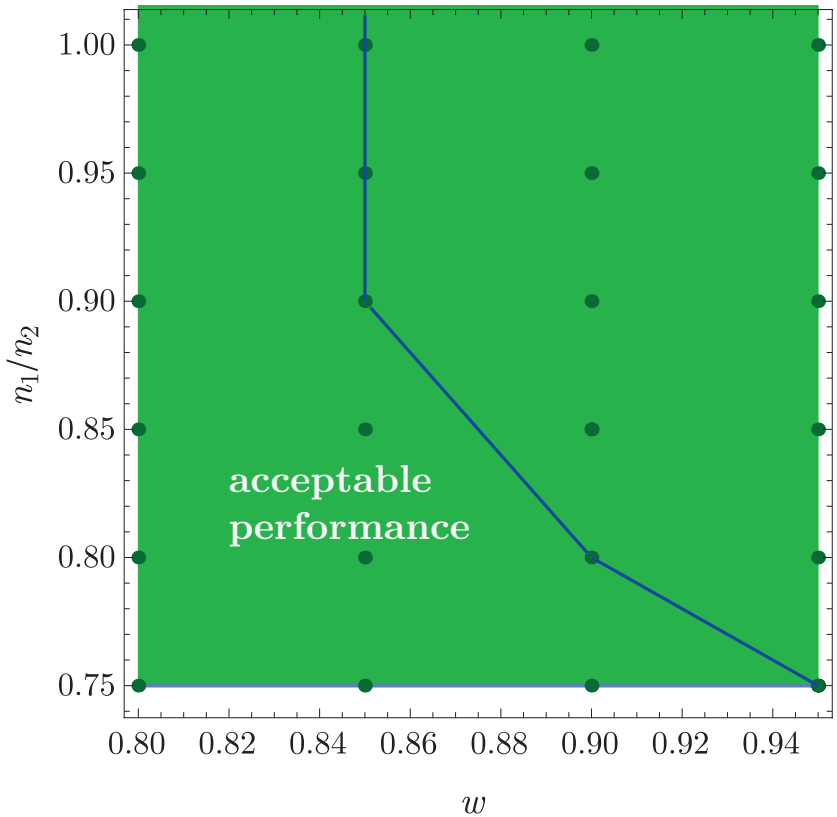
Algorithm performance (Approach B)

Detection of anomalous nodes



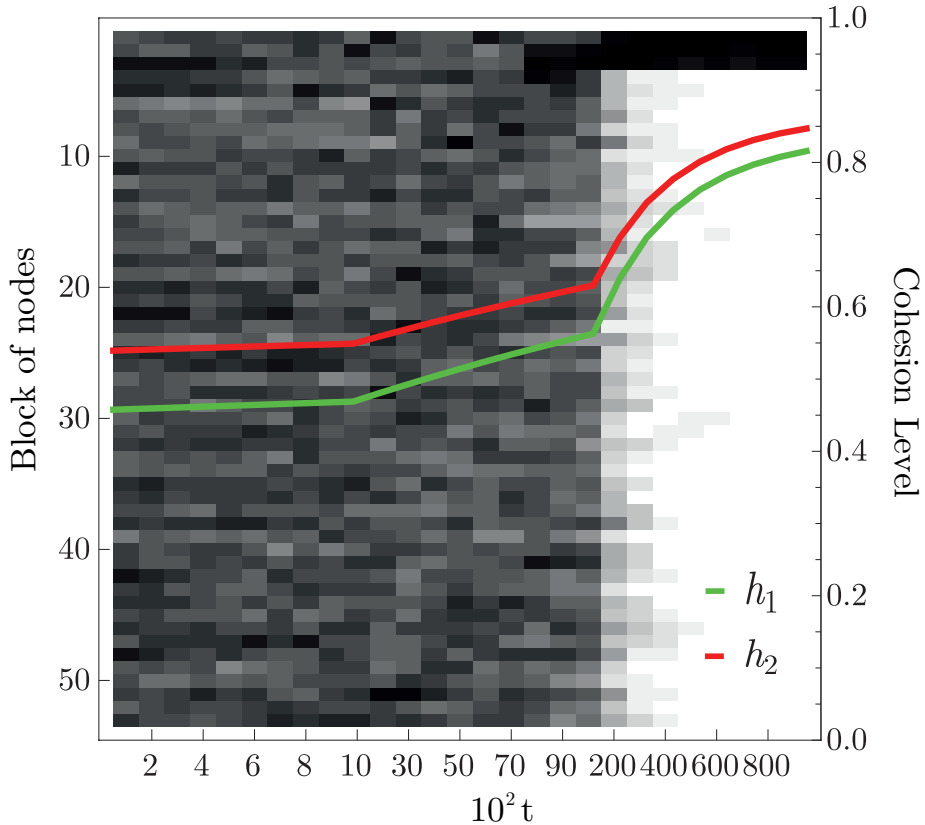
Algorithm performance (Approach B)

Area of acceptable performance



Dynamic detection (Approach B)

Approach B - Identifying suspects

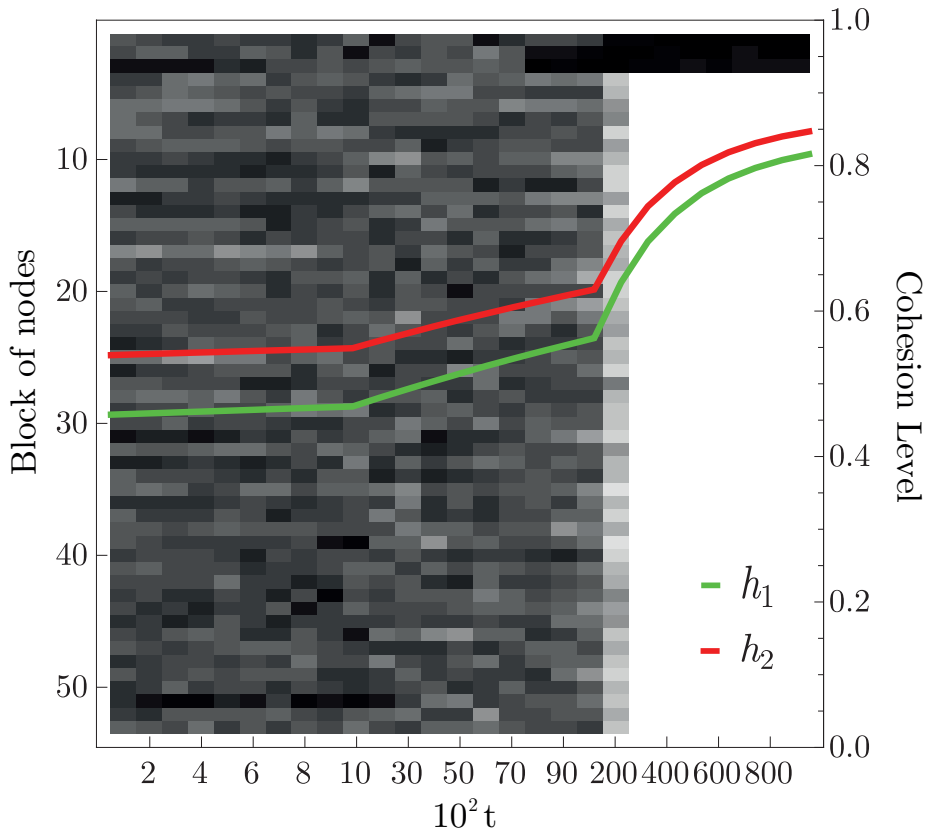


acceptable performance when $h_1, h_2 > 0.72$

$\mathrm{t} \rightarrow 500 \times 10^2$

Dynamic detection (Approach B)

Approach B - Detecting anomalous nodes

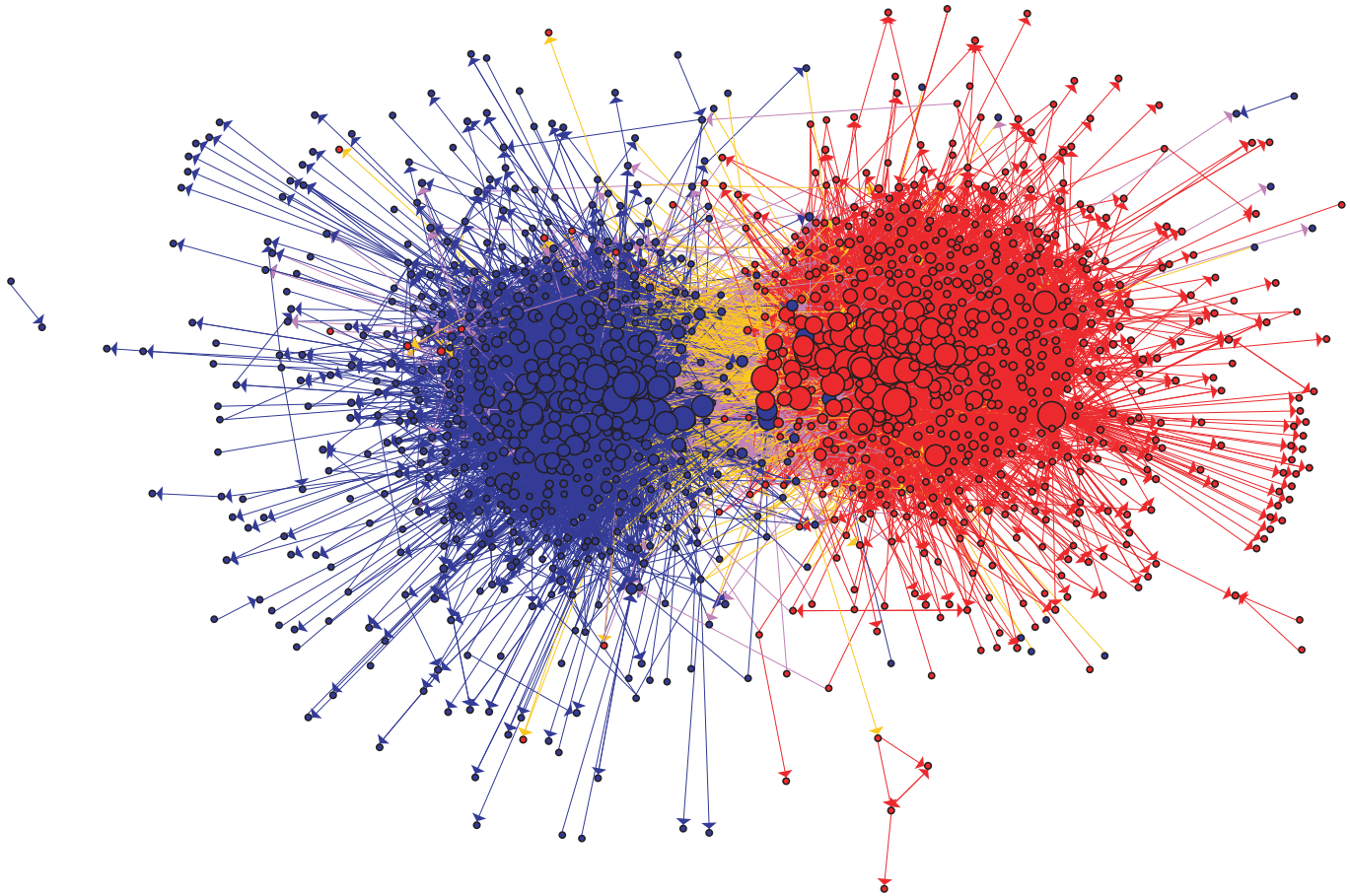


acceptable performance when $h_1, h_2 > 0.64$

$$\text{t} \rightarrow 250 \times 10^2$$

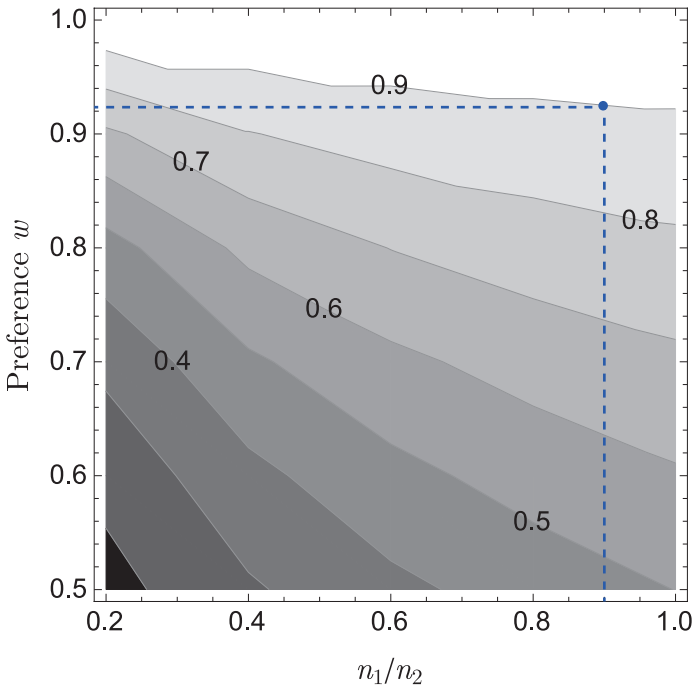
Empirical Network-Static Detection

Political blogosphere network (polblogs) preceding the 2004 U.S. presidential election

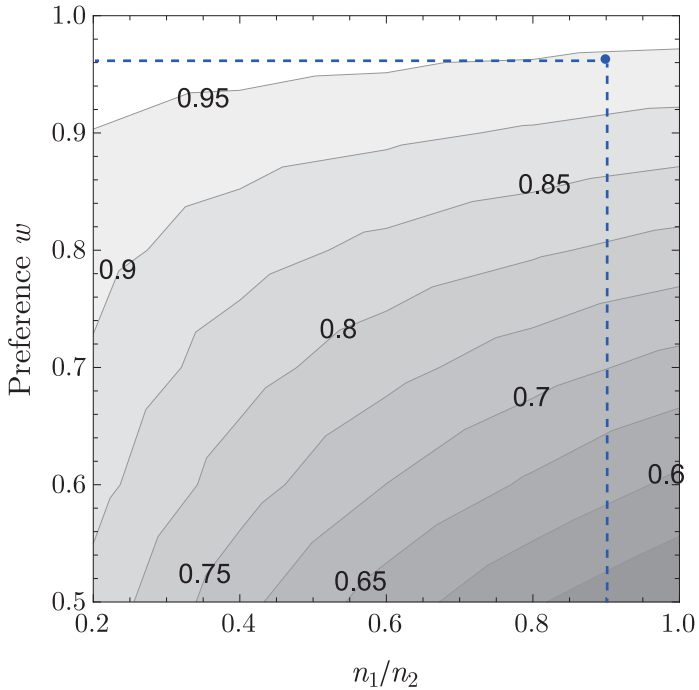


Empirical Network-Static Detection

- $n_1/n_2 = 0.9$
- $h_1 = 0.91,\ h_2 = 0.95$

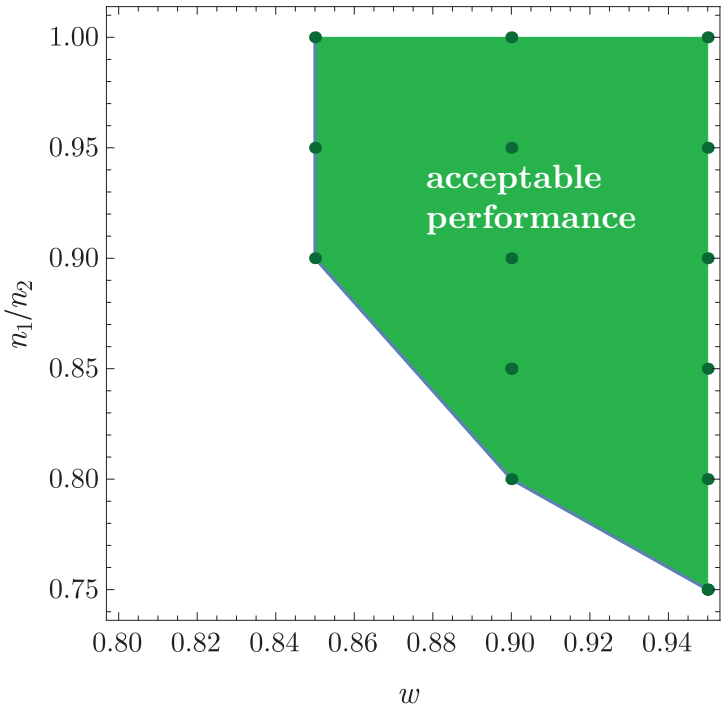


minority group

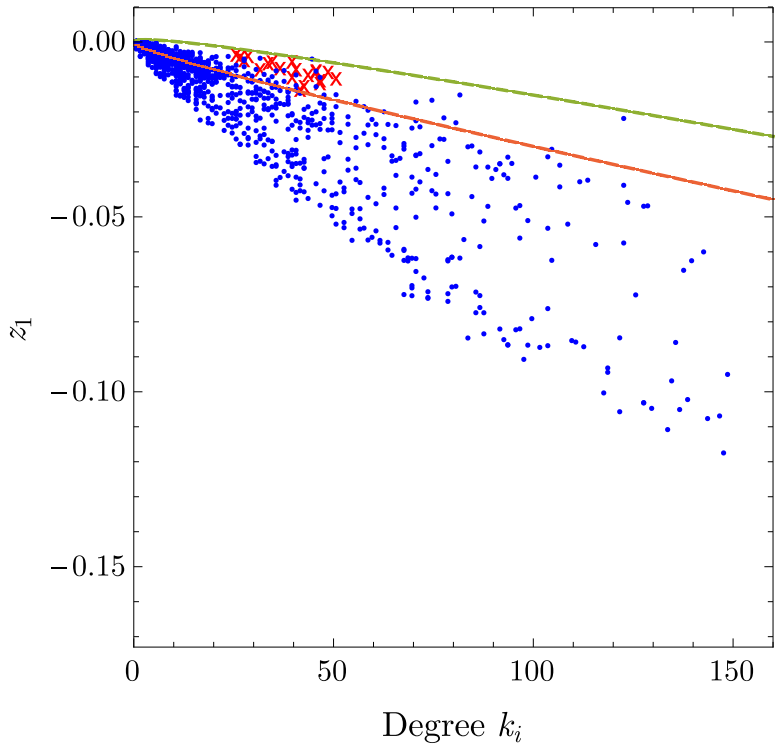


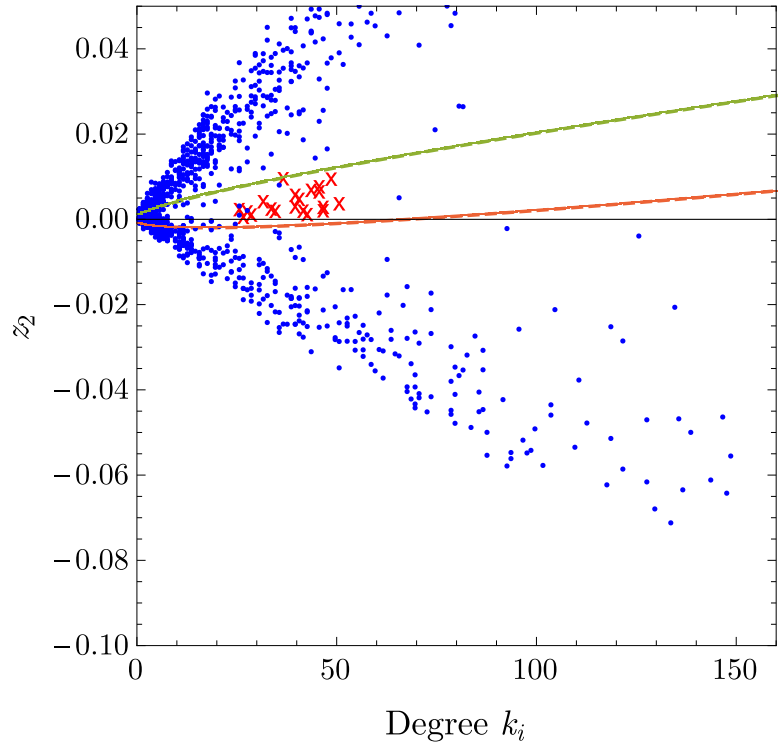
majority group

$w \approx 0.95$



	e_1	e_2
Approach B - Step1	0.91	0.95
Approach B	0	0.95
Approach A - Step1	0.91	1.0
Approach A	0	1.0

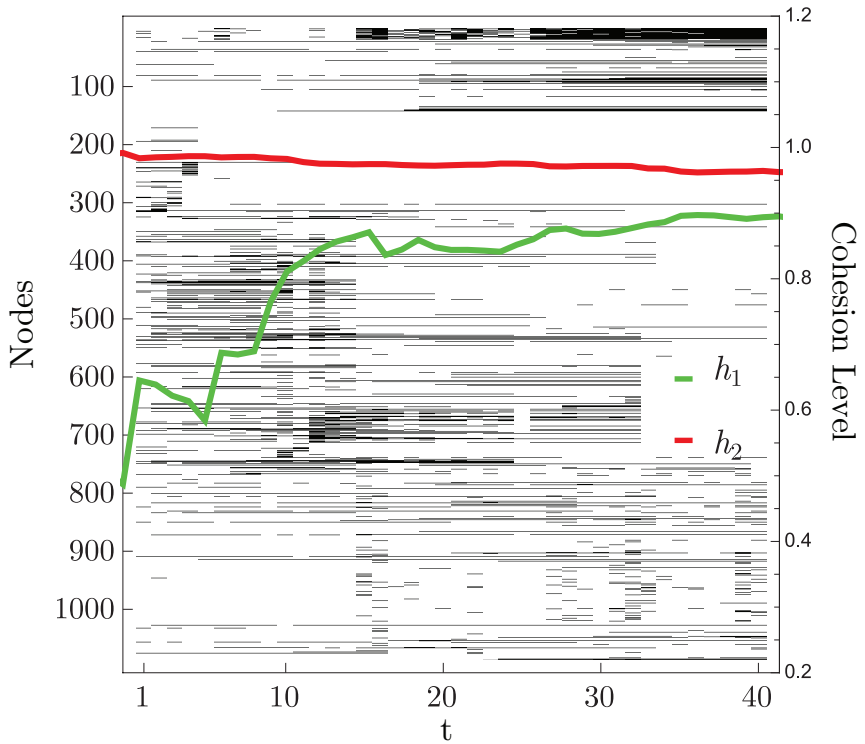




Proposed approach tends to identify nodes with a low degree as suspe

Empirical Network Dynamic Detection

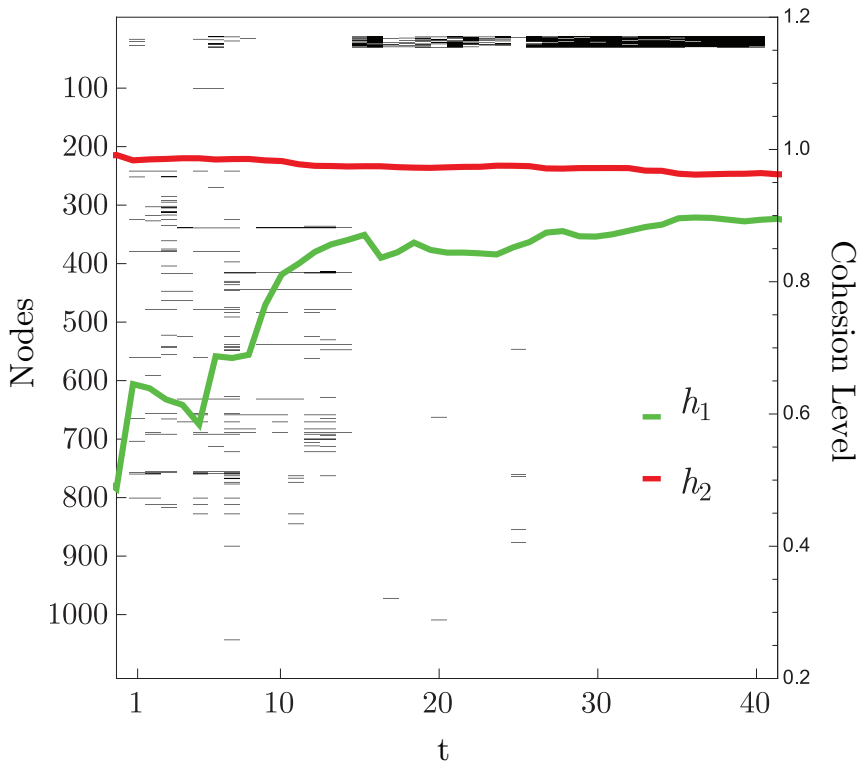
Approach B - Identifying suspects



acceptable performance is never achieved

Empirical Network Dynamic Detection

Approach B - Detecting anomalous nodes



acceptable performance when the cohesion levels $h_1, h_2 > 0.85$

Conclusions

- The model reaches stationary cohesion and modularity levels for regular nodes.

- The proposed approach increases the area of acceptable performance.

- The model serves as an analytical framework to establish detection thresholds for acceptable performance.

- The cohesion indices are a key criterion to determine the effectiveness of detection algorithms.