Exploring the Impact of Trolls on Activity Dynamics in Real-World Collaboration Networks

Philipp Koncar [1], Simon Walk [1, 2], Denis Helic [1], Markus Strohmaier [3]

April 3, 2017

[1]Graz University of Technology
[2]Stanford University
[3]Gesis & University of Koblenz-Landau

Motivation

Online Collaboration Networks

Highly dynamic and depending on sustainable user activity.

The Problem

Online collaboration networks are often targeted by trolls!
Consequences are hard to foresee.

Trolls

Users that intentionally contribute detrimental content to harm the networks.

The Goal

First step towards tools that help website owners to estimate and ease the impact of trolls.

Activity Dynamics Model [1]

Concept

Based on the principles of dynamical systems on networks.

Networks Represented by Graphs

Users represented by nodes.
Collaboration represented by edges.

Two Opposing Forces

I) Users tend to lose interest to contribute.
II) Users tend to react to actions of their peers.

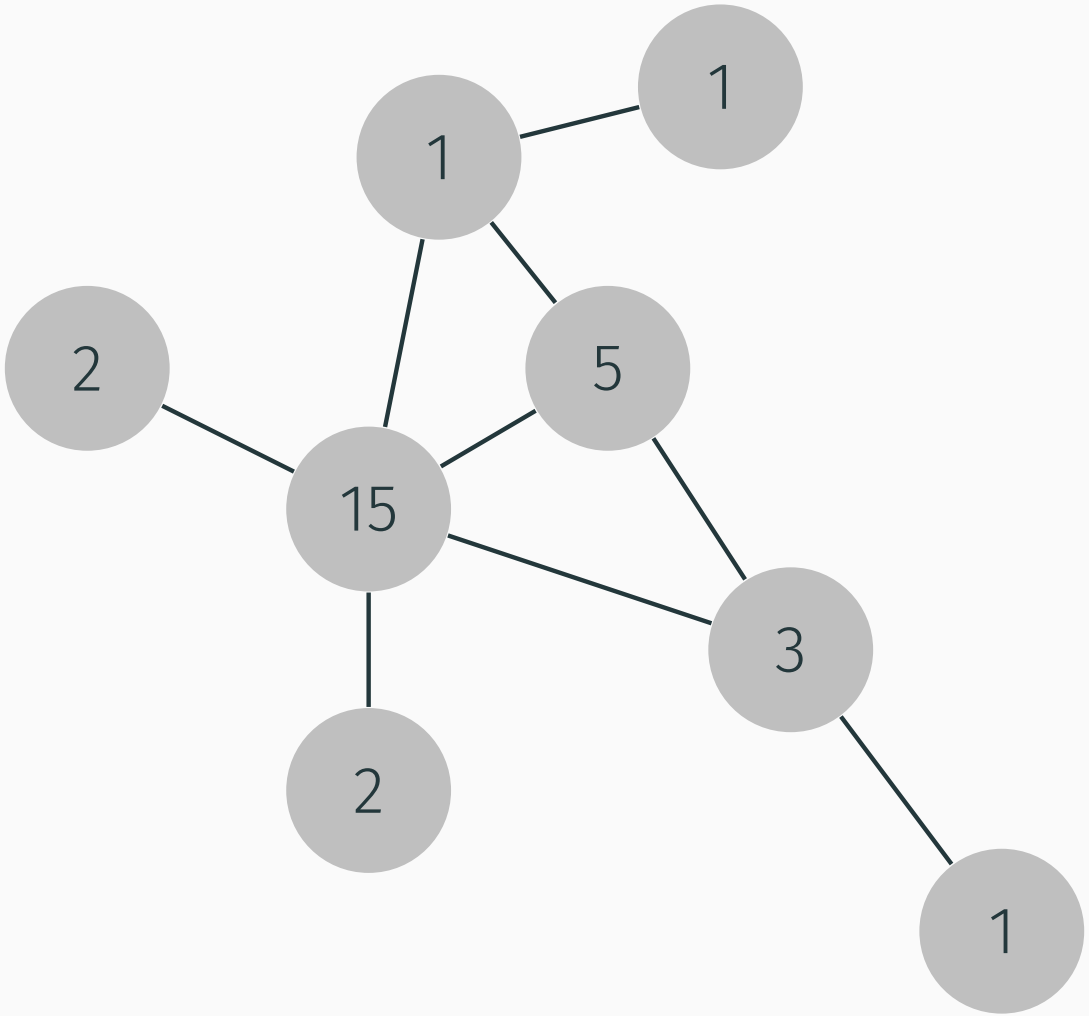
One Single Parameter λ/μ

Fitted by using a least-squares approach and empirical data.

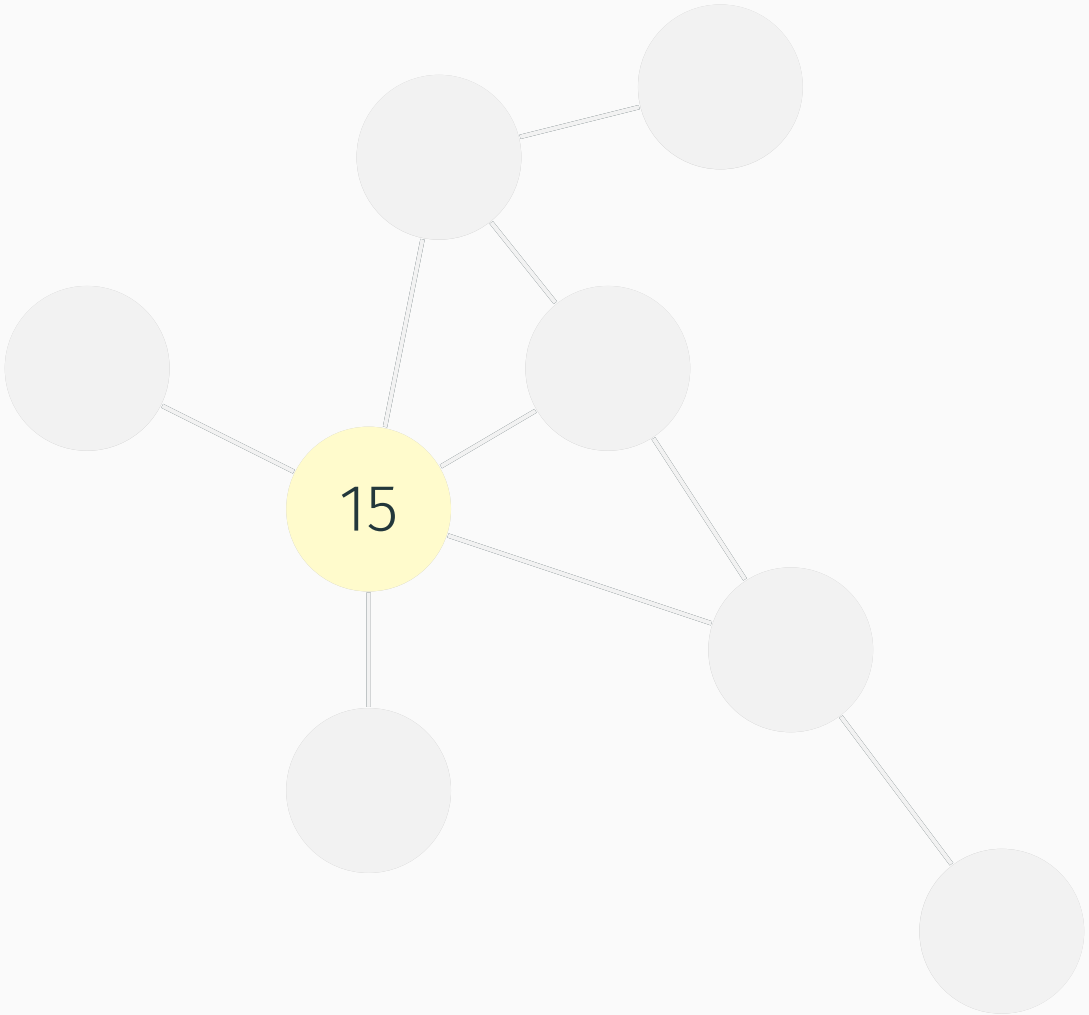
---

[1] Walk, S., Helic, D., Geigl, F., & Strohmaier, M. (2016). Activity dynamics in collaboration networks. ACM Transactions on the Web (TWEB), 10(2), 11.

Experiments

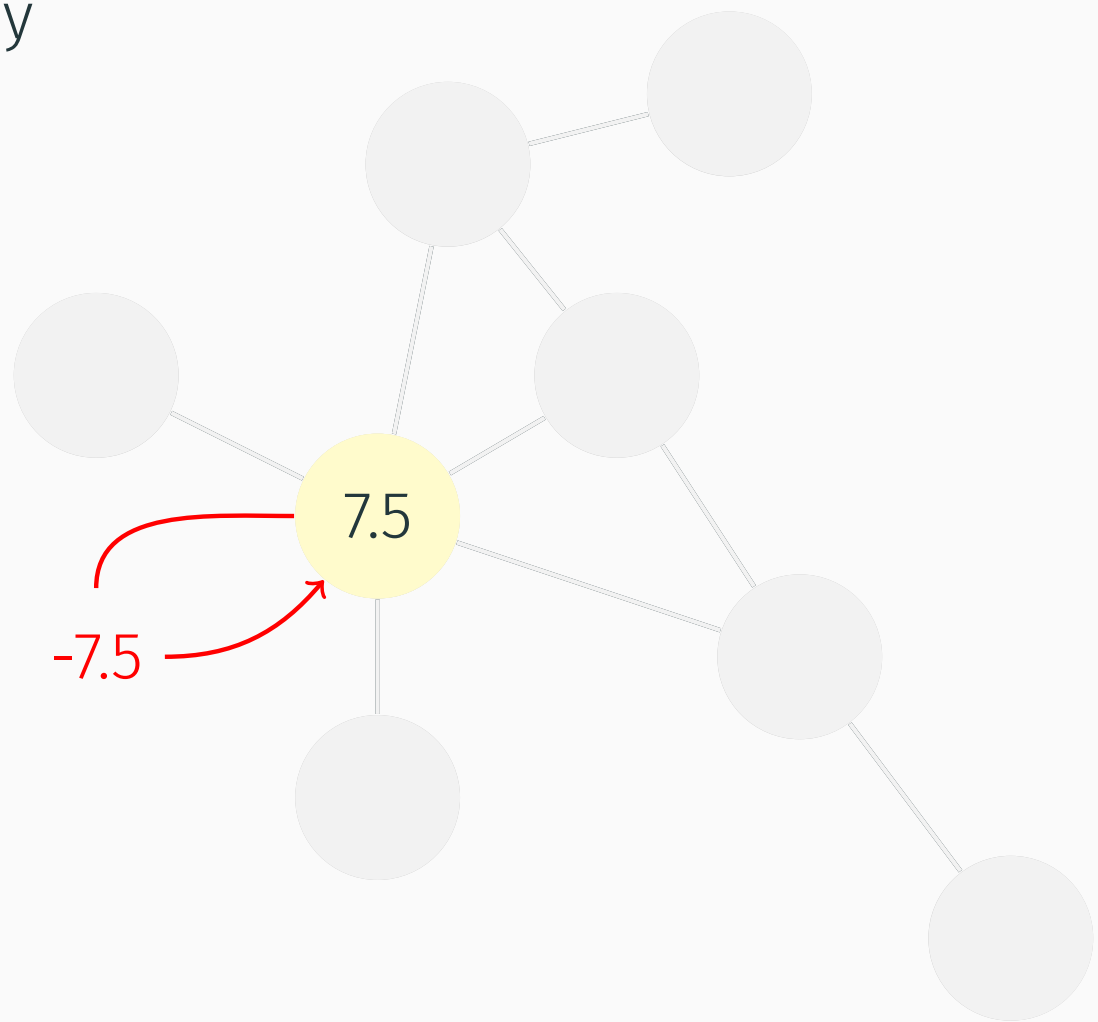


Experiments



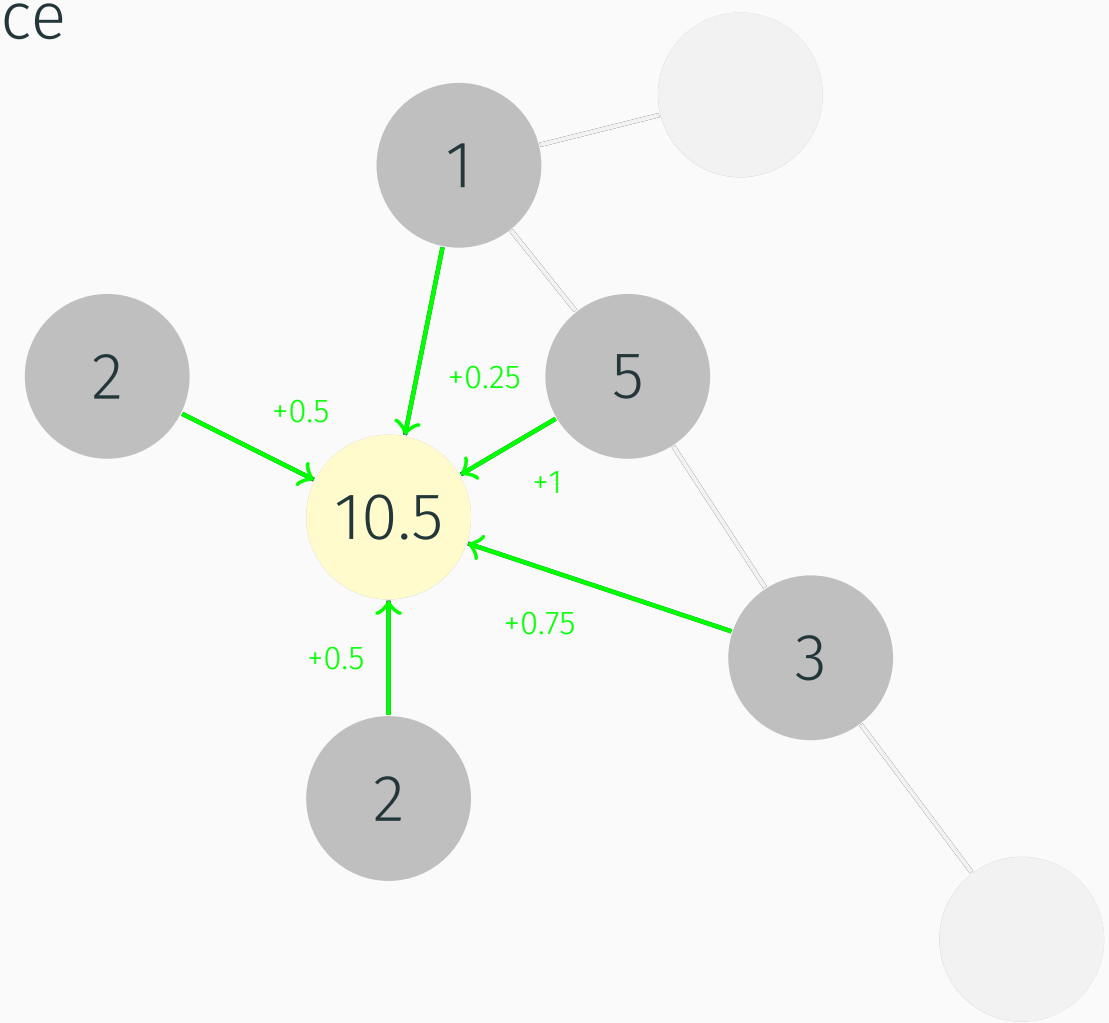
Experiments

Activity Decay



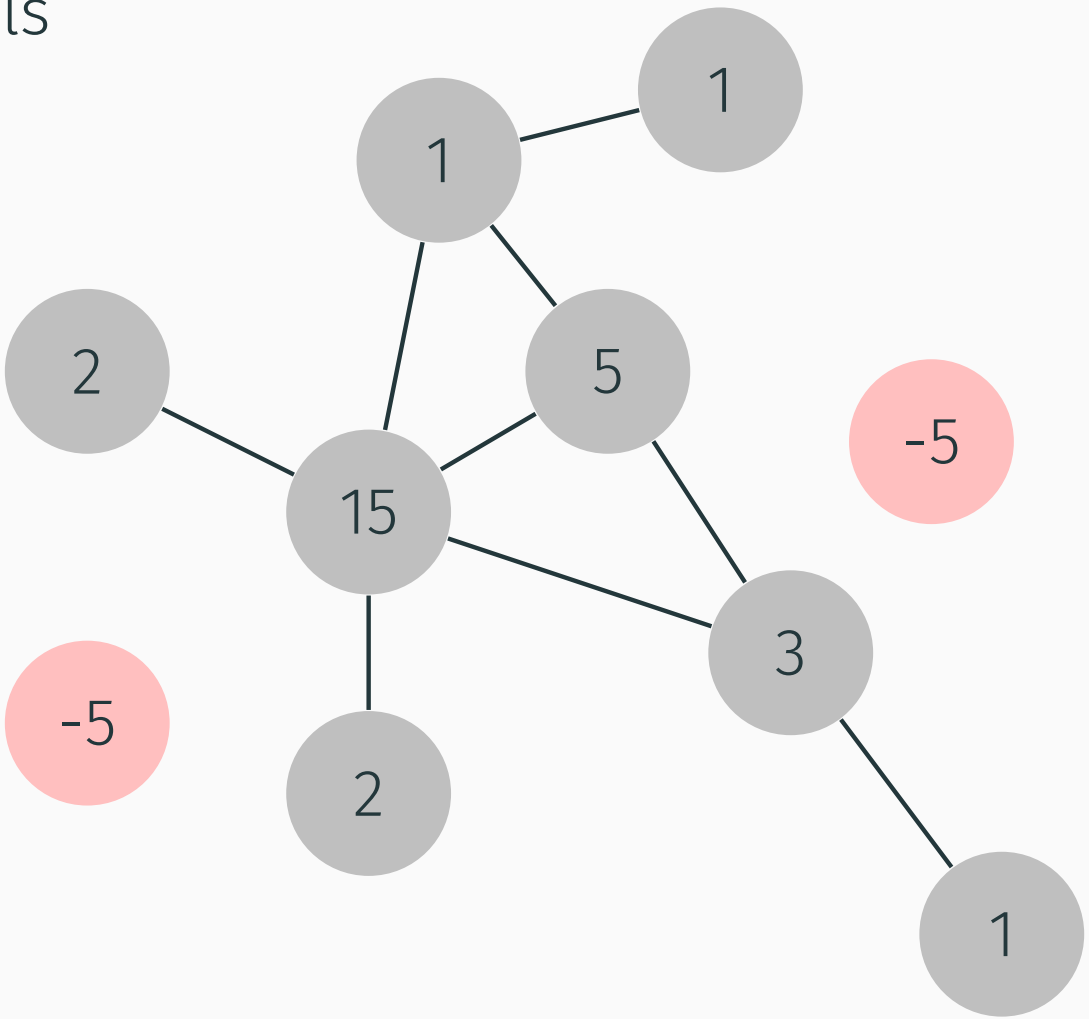
Experiments

Peer Influence



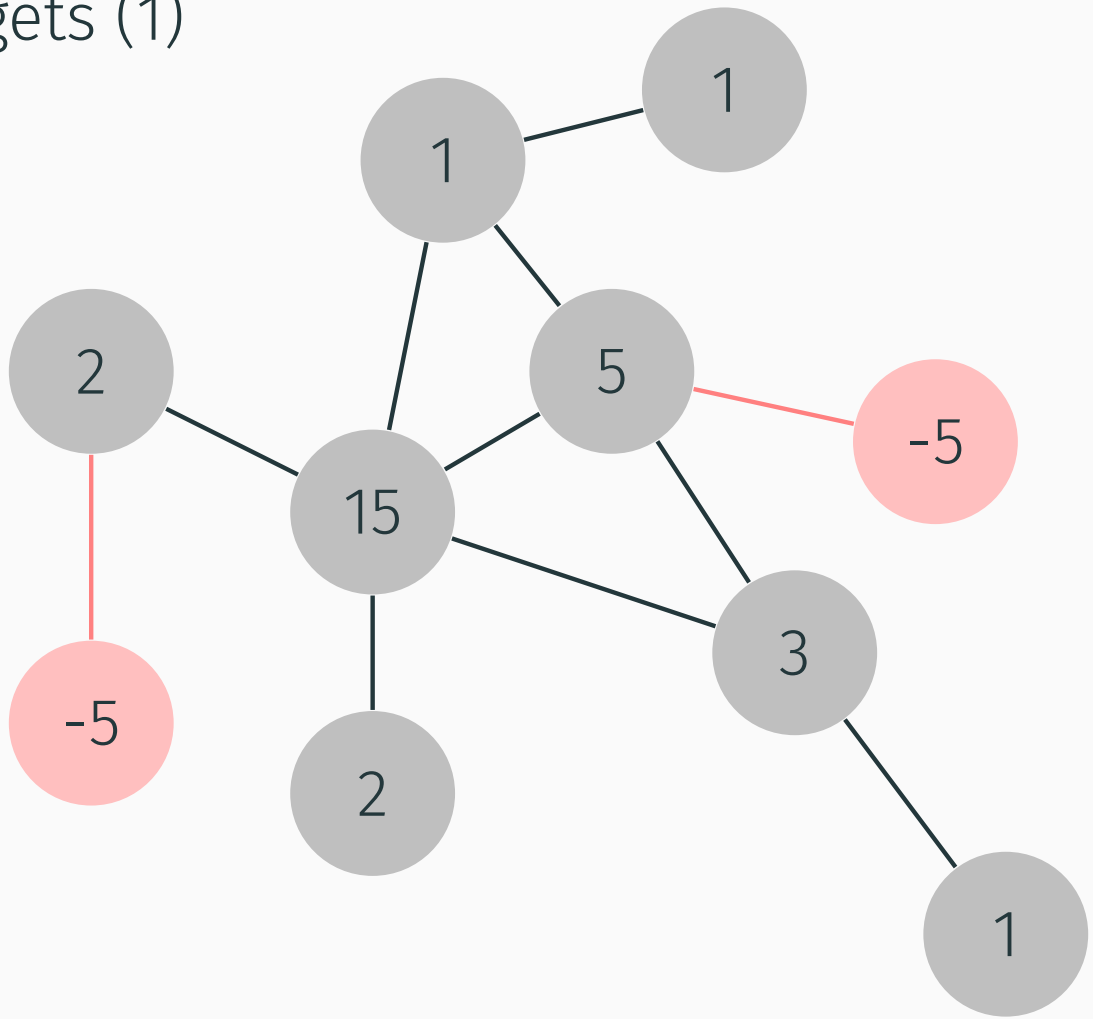
Experiments

Adding 2 Trolls



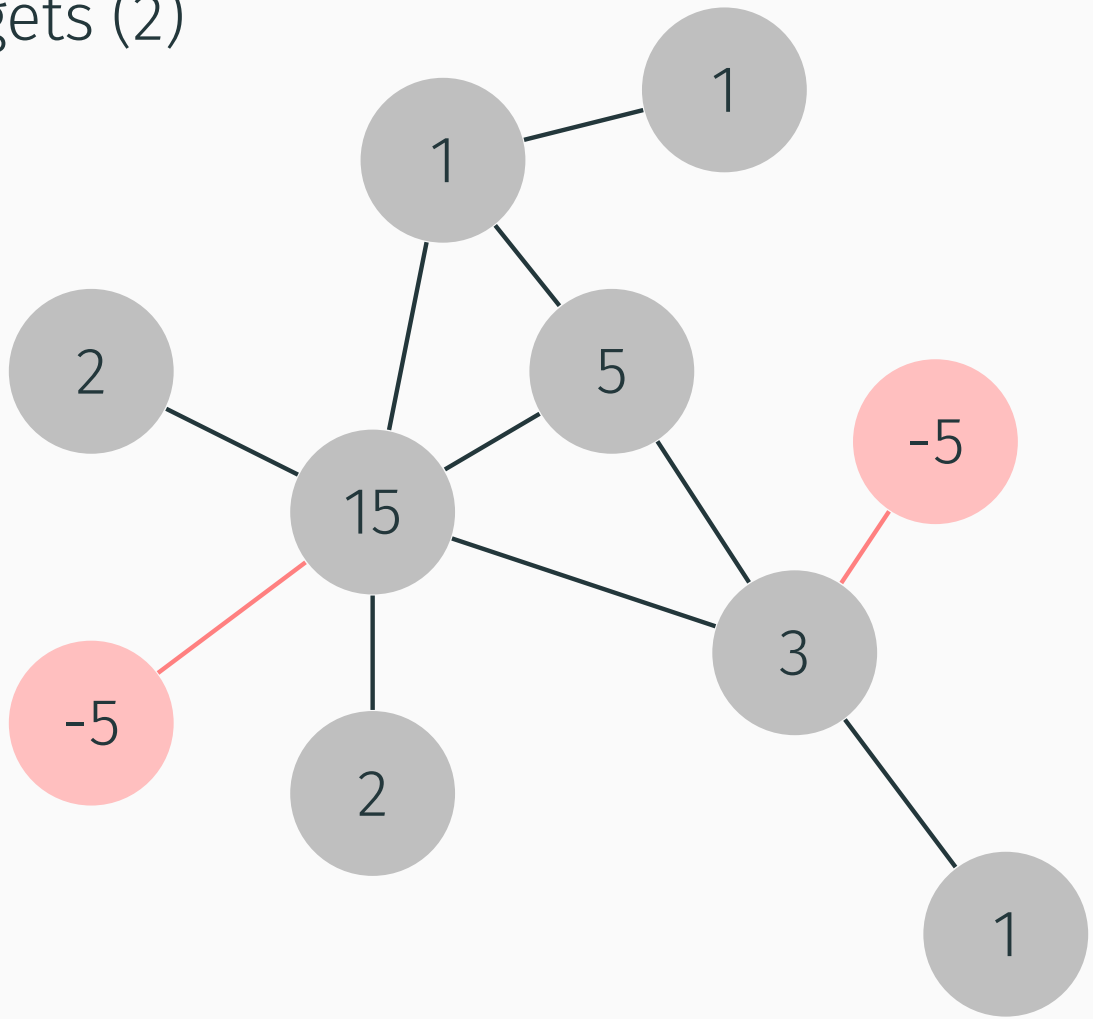
Experiments

Random Targets (1)



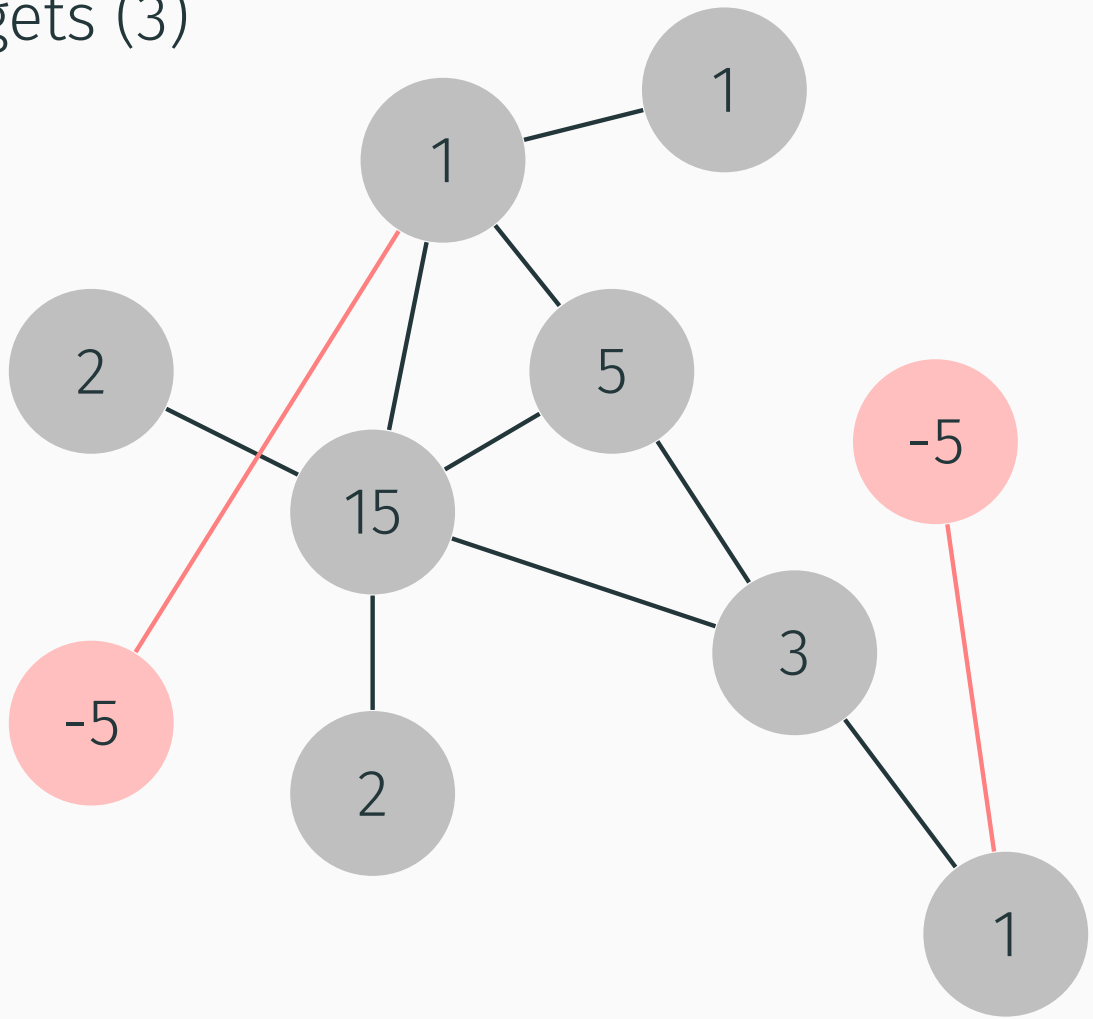
Experiments

Random Targets (2)



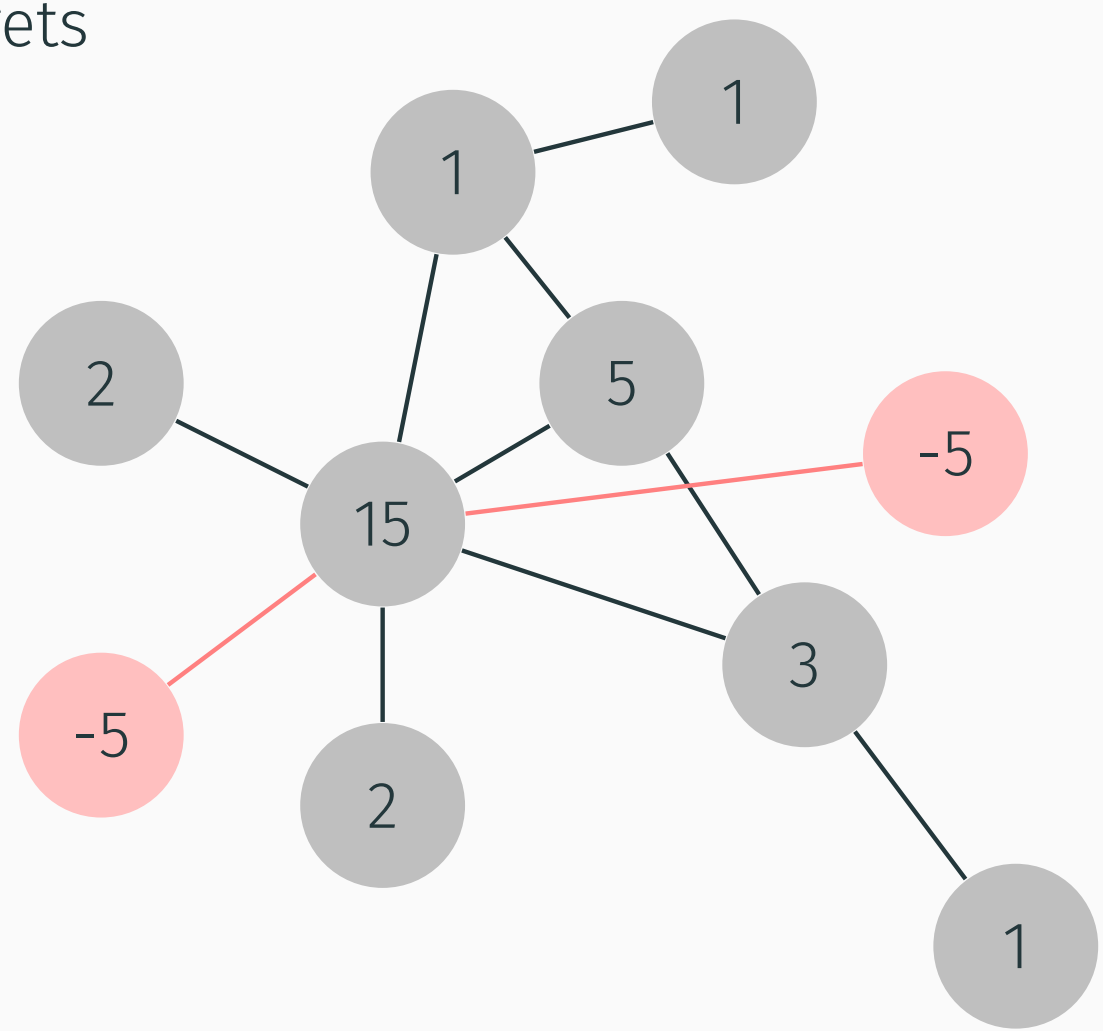
Experiments

Random Targets (3)



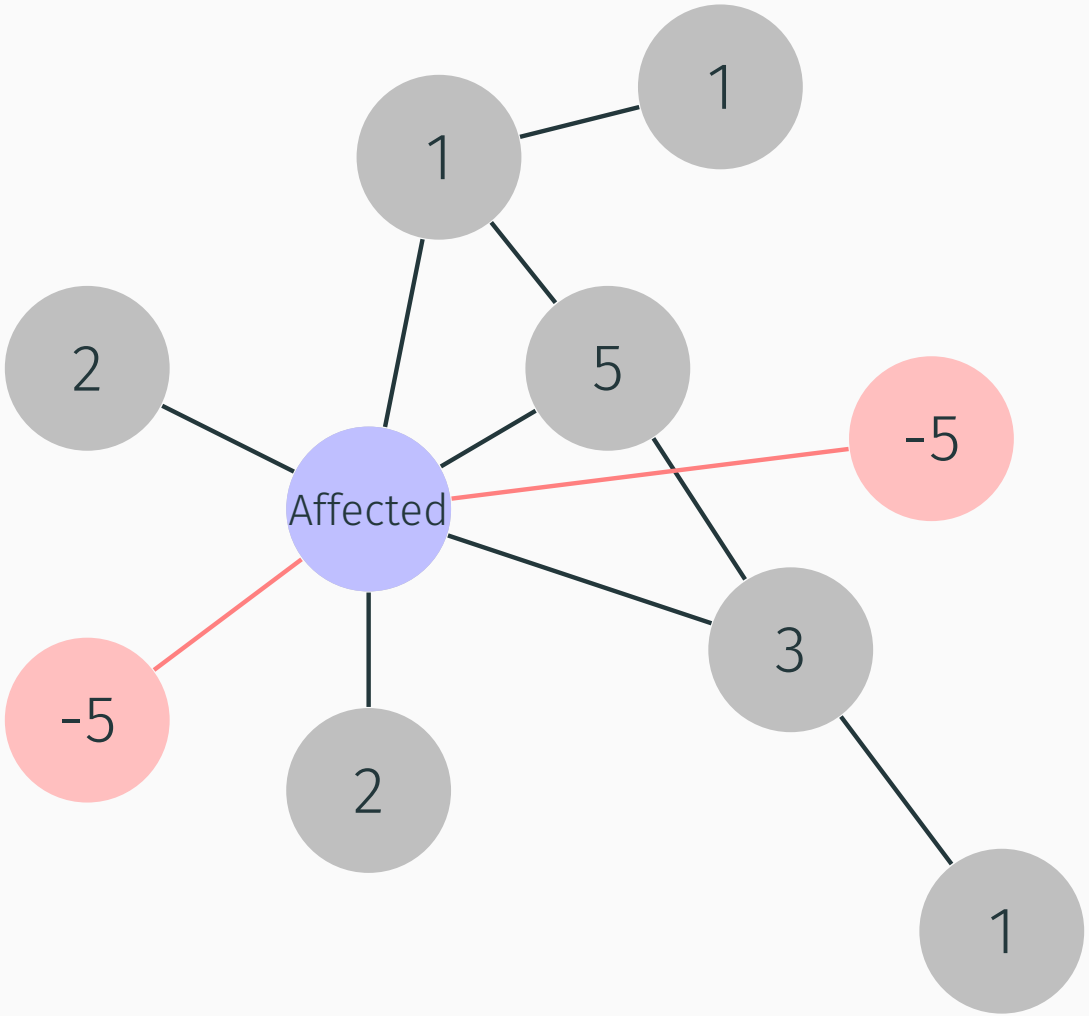
Experiments

Specific Targets



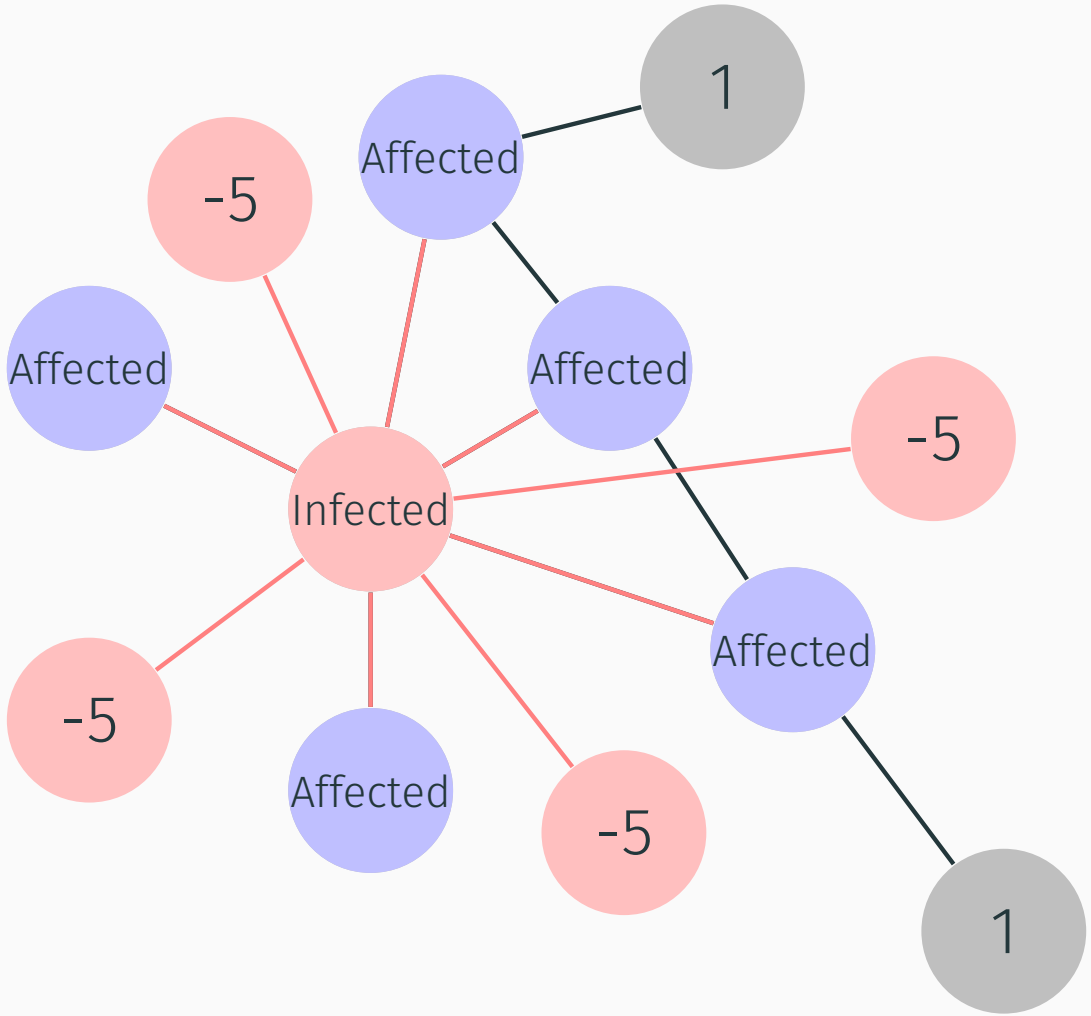
Experiments

2 Trolls



Experiments

4 Trolls



Experiments

Adding Trolls

Adding nodes with initial activity set to **-5**, representing detrimental content.

Two Different Strategies

I) random: randomly select affected users
II) informed: specifically target high degree users

Affected Users

Users that were exposed to negative activity.

Infected Users

Users that started to contribute negative activity by themselves.

Data sets

Dataset	StackExchange		Semantic MediaWikis		Subreddits	
	Bitcoin	English	DotaWiki	NeuroLex	r/Austria	r/StarWars
Users	1,346	9,191	233	114	1,454	31,121
Edges	5,653	96,982	725	384	8,234	208,881
Posts & Replies	14,242	181,033	17,197	36,461	16,329	305,181
Mean Degree	8	21	6	7	11	13
Median Degree	3	6	3	3	5	5
Start	02/16/2014	02/16/2014	12/07/2008	11/18/2012	12/08/2013	12/08/2013
End	03/08/2015	03/08/2015	11/27/2009	12/08/2013	12/28/2014	12/28/2014
# Weeks	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$
NRMSE	0.13	0.22	0.25	0.34	0.15	0.14

Data sets

Dataset	StackExchange		Semantic MediaWikis		Subreddits	
	Bitcoin	English	DotaWiki	NeuroLex	r/Austria	r/StarWars
Users	1,346	9,191	233	114	1,454	31,121
Edges	5,653	96,982	725	384	8,234	208,881
Posts & Replies	14,242	181,033	17,197	36,461	16,329	305,181
Mean Degree	8	21	6	7	11	13
Median Degree	3	6	3	3	5	5
Start	02/16/2014	02/16/2014	12/07/2008	11/18/2012	12/08/2013	12/08/2013
End	03/08/2015	03/08/2015	11/27/2009	12/08/2013	12/28/2014	12/28/2014
# Weeks	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$
NRMSE	0.13	0.22	0.25	0.34	0.15	0.14

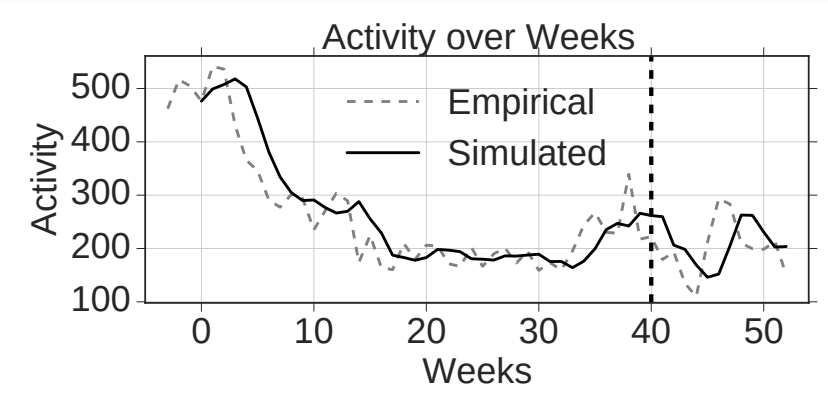
Data sets

Dataset	StackExchange		Semantic MediaWikis		Subreddits	
	Bitcoin	English	DotaWiki	NeuroLex	r/Austria	r/StarWars
Users	1,346	9,191	233	114	1,454	31,121
Edges	5,653	96,982	725	384	8,234	208,881
Posts & Replies	14,242	181,033	17,197	36,461	16,329	305,181
Mean Degree	8	21	6	7	11	13
Median Degree	3	6	3	3	5	5
Start	02/16/2014	02/16/2014	12/07/2008	11/18/2012	12/08/2013	12/08/2013
End	03/08/2015	03/08/2015	11/27/2009	12/08/2013	12/28/2014	12/28/2014
# Weeks	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$
NRMSE	0.13	0.22	0.25	0.34	0.15	0.14

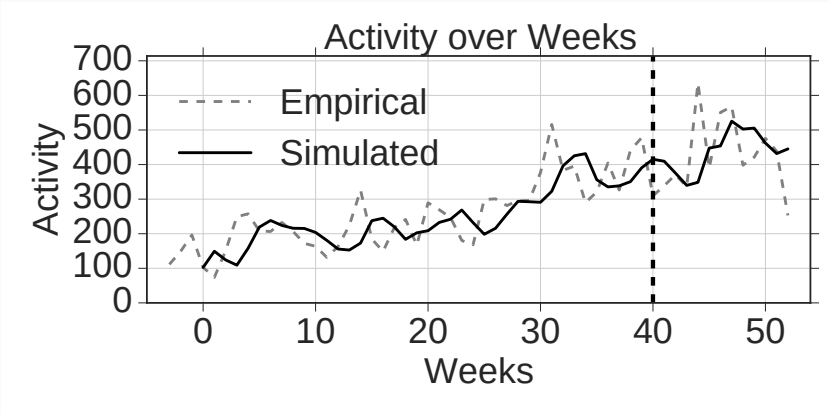
Data sets

Dataset	StackExchange		Semantic MediaWikis		Subreddits	
	Bitcoin	English	DotaWiki	NeuroLex	r/Austria	r/StarWars
Users	1,346	9,191	233	114	1,454	31,121
Edges	5,653	96,982	725	384	8,234	208,881
Posts & Replies	14,242	181,033	17,197	36,461	16,329	305,181
Mean Degree	8	21	6	7	11	13
Median Degree	3	6	3	3	5	5
Start	02/16/2014	02/16/2014	12/07/2008	11/18/2012	12/08/2013	12/08/2013
End	03/08/2015	03/08/2015	11/27/2009	12/08/2013	12/28/2014	12/28/2014
# Weeks	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$
NRMSE	0.13	0.22	0.25	0.34	0.15	0.14

Activity Dynamics Simulations



(a) BitcoinStackExchange



(b) r/Austria

Adding Trolls (Part I)

Motivation

How will the networks react to added trolls?

Number of Trolls

0.25%, 0.50% and 1.00% of existing users

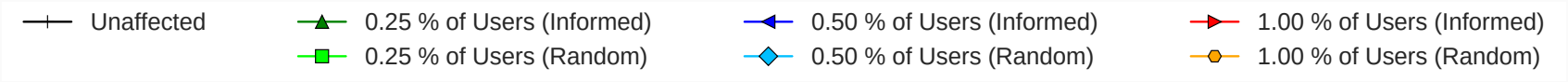
Number of Connections

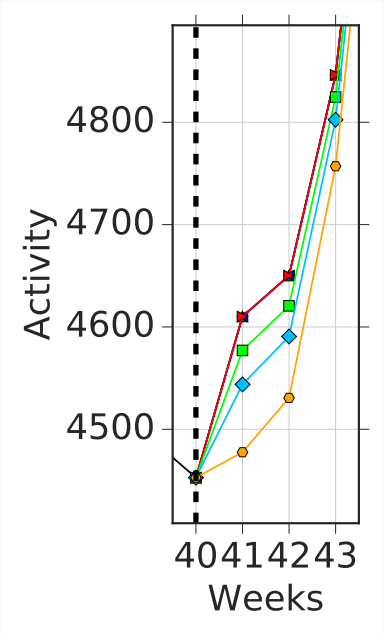
Average Degrees

Observed Changes

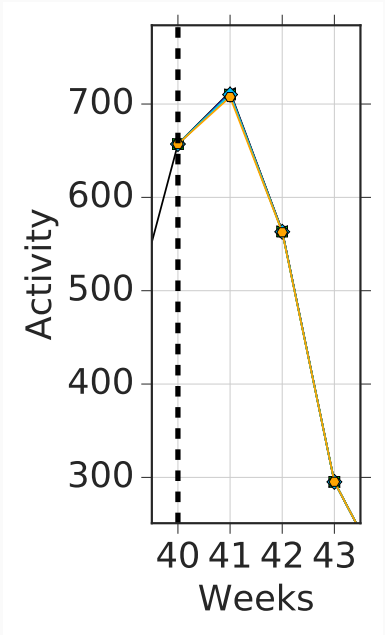
Overall network activity per week

Adding Trolls (Part I)



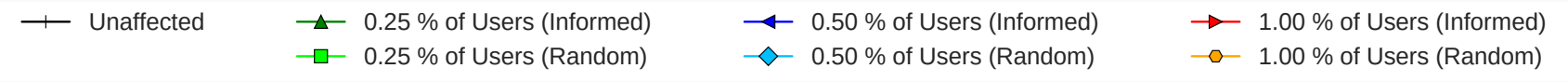


r/StarWars



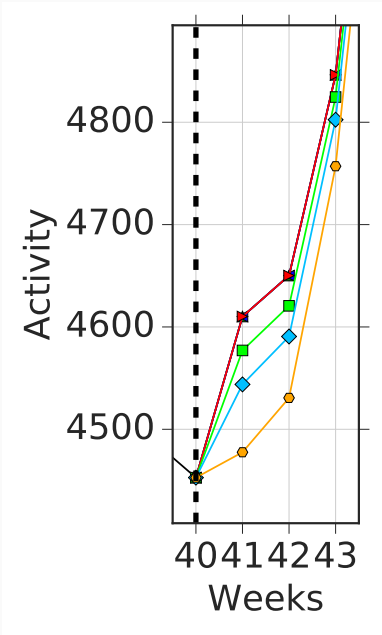
DotaWiki

Adding Trolls (Part I)

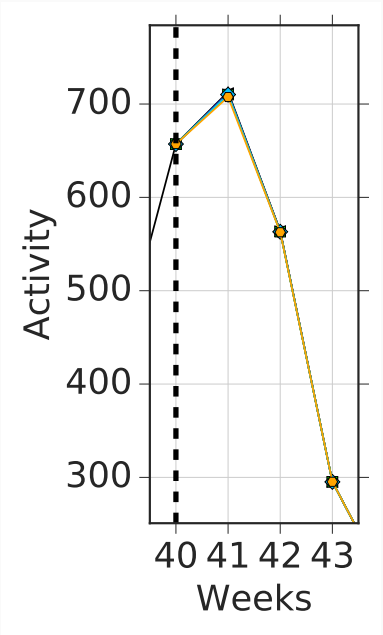


Informed

Minimal impact =
High peer influence
High activity levels

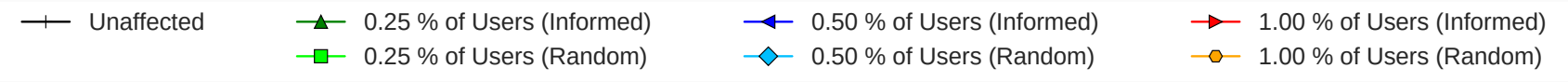


r/StarWars



DotaWiki

Adding Trolls (Part I)

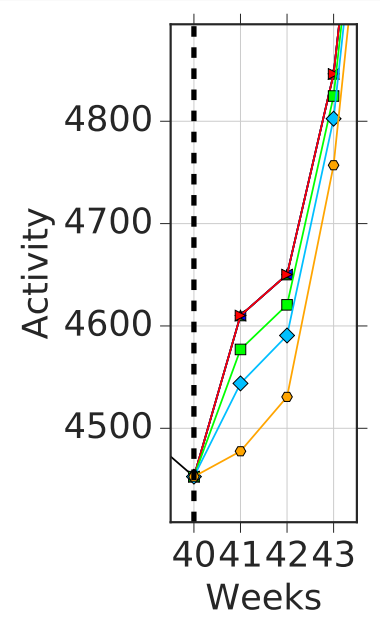


Informed

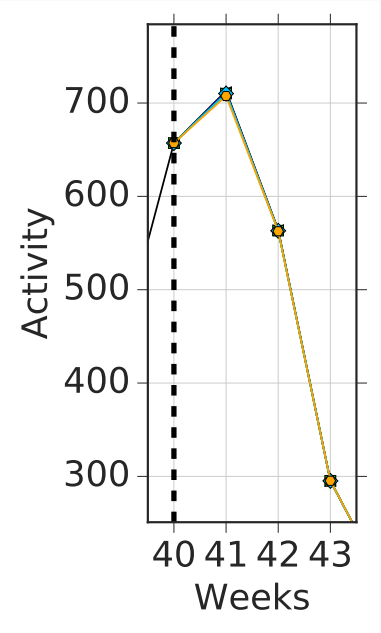
Minimal impact =
High peer influence
High activity levels

Random

High impact =
Low peer influence
Low activity levels

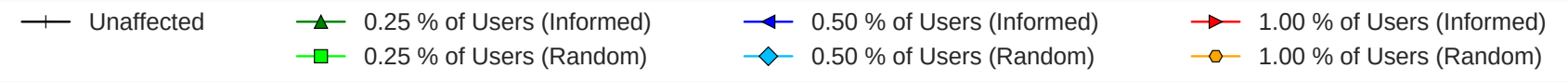


r/StarWars



DotaWiki

Adding Trolls (Part I)

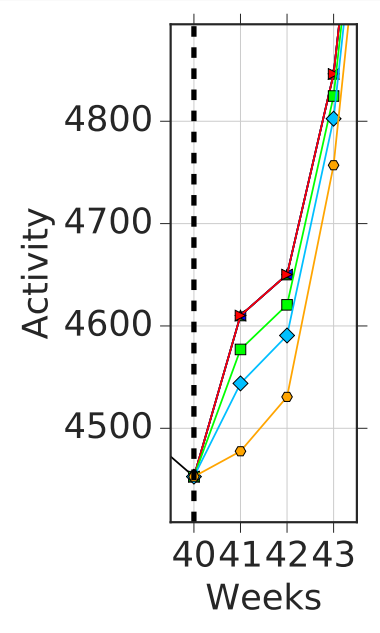


Informed

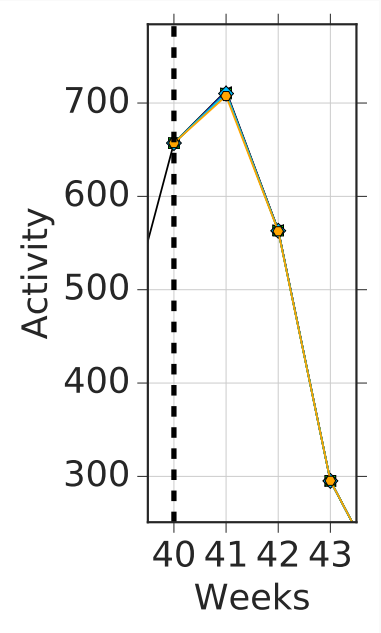
Minimal impact =
High peer influence
High activity levels

Random

High impact =
Low peer influence
Low activity levels



r/StarWars

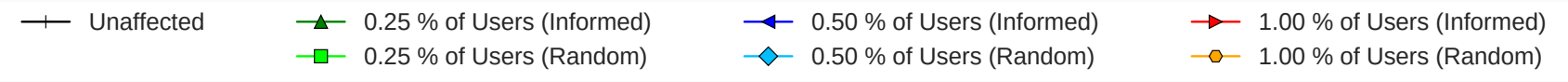


DotaWiki

Informed

Minimal impact =
High peer influence
High activity levels

Adding Trolls (Part I)

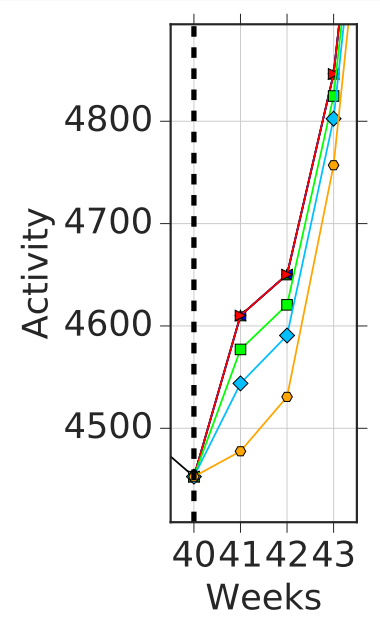


Informed

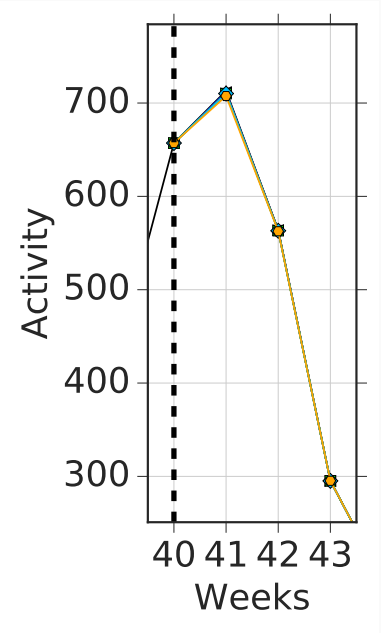
Minimal impact =
High peer influence
High activity levels

Random

High impact =
Low peer influence
Low activity levels



r/StarWars



DotaWiki

Informed

Minimal impact =
High peer influence
High activity levels

Random

Minimal impact =
High peer influence
High activity levels

Adding Trolls (Part II)

Motivation

How much trolls need to be added to collapse the whole network?

Number of Trolls

1.00% to 5.00% incremented by 0.10%

Number of Connections

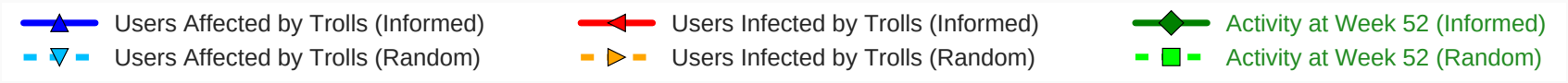
Average Degrees

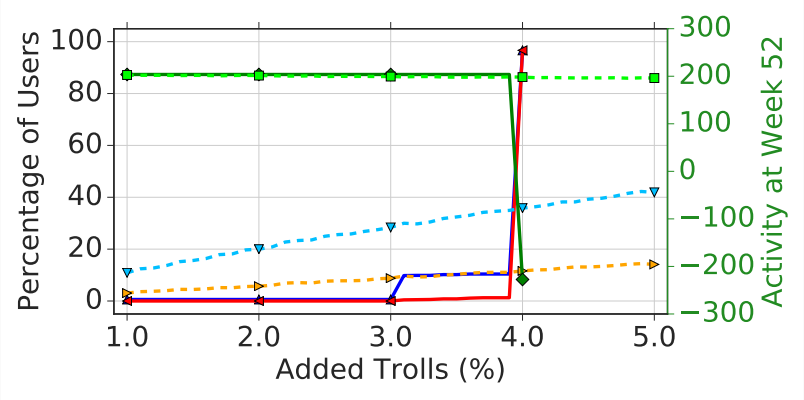
Observed Changes

Overall network activity
users affected by trolls
users infected by trolls

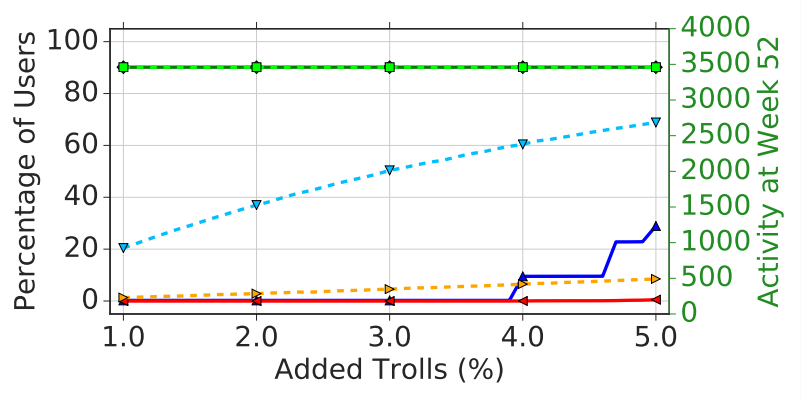
} at the end of simulations.

Adding Trolls (Part II)



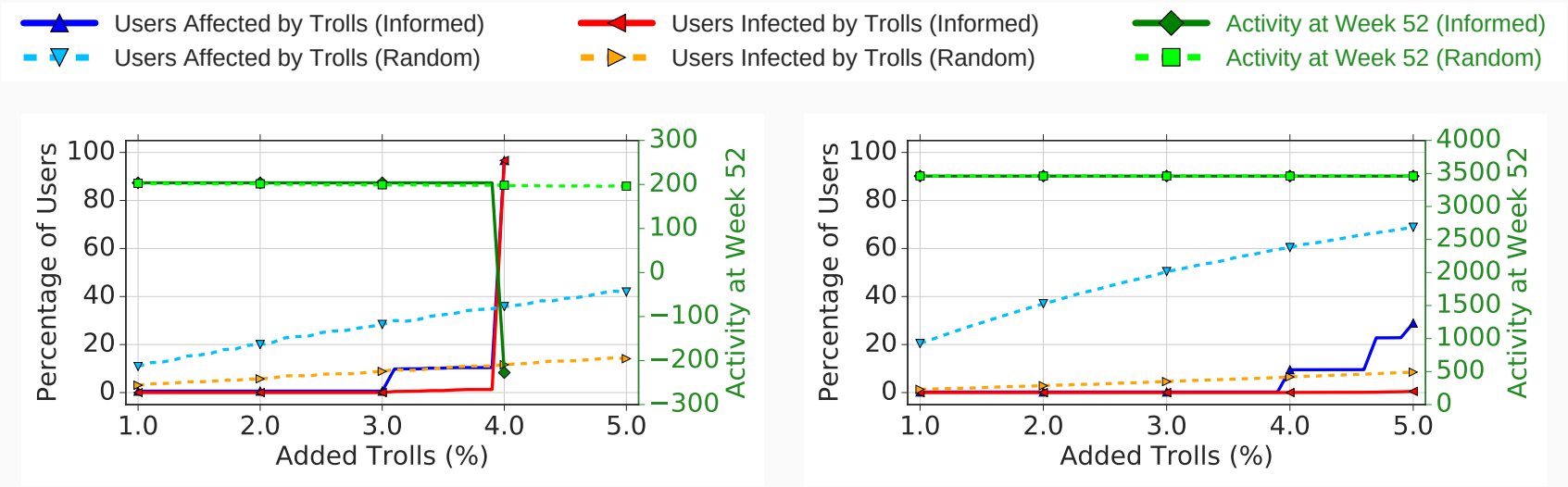


BitcoinStackExchange



EnglishStackExchange

Adding Trolls (Part II)



BitcoinStackExchange

EnglishStackExchange

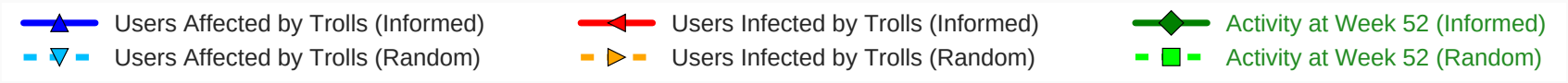
Informed

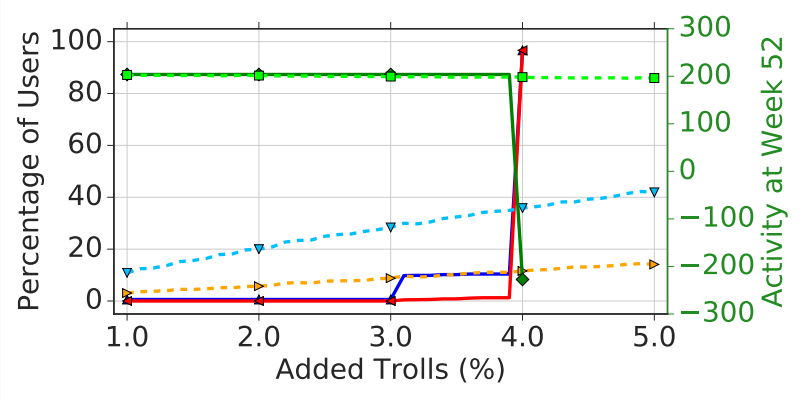
Network abruptly collapsed

Random

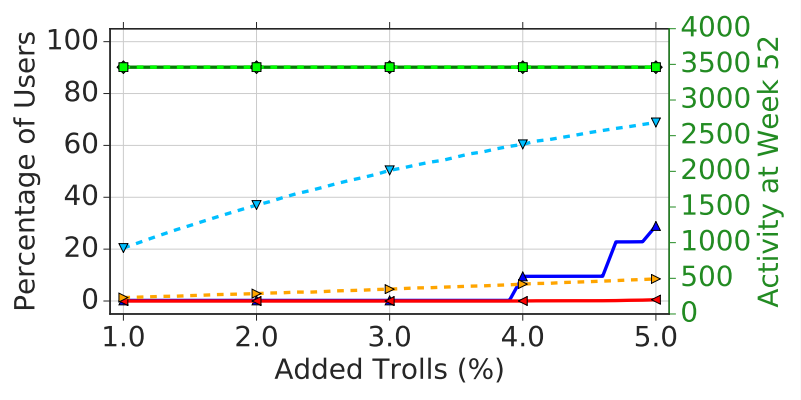
Almost 50% of users affected
No impact on activity

Adding Trolls (Part II)





BitcoinStackExchange



EnglishStackExchange

Informed

Network abruptly collapsed

Random

Almost 50% of users affected
No impact on activity

Informed

No impact on activity

Random

70% of users affected
No impact on activity

Increasing Trolls' Exposure

Motivation

What happens if trolls reach out to more existing users?

Number of Trolls

1.00% of existing users

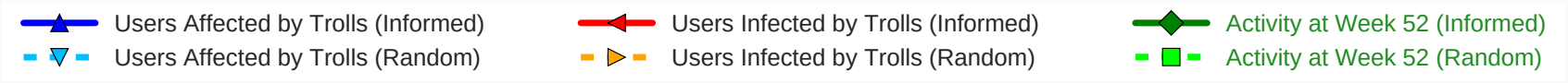
Number of Connections

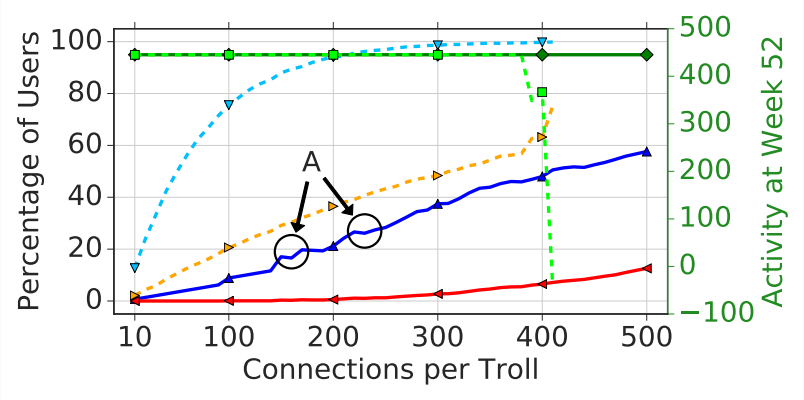
10 to 500 incremented by 10

Observed Changes

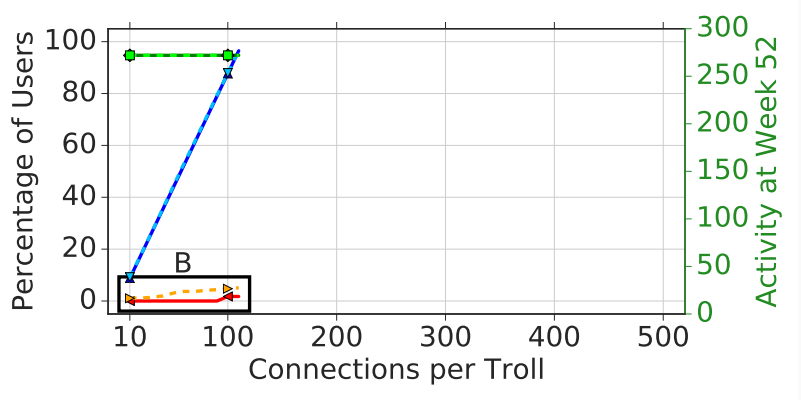
Overall network activity
users affected by trolls
users infected by trolls } at the end of simulations.

Increasing Trolls' Exposure



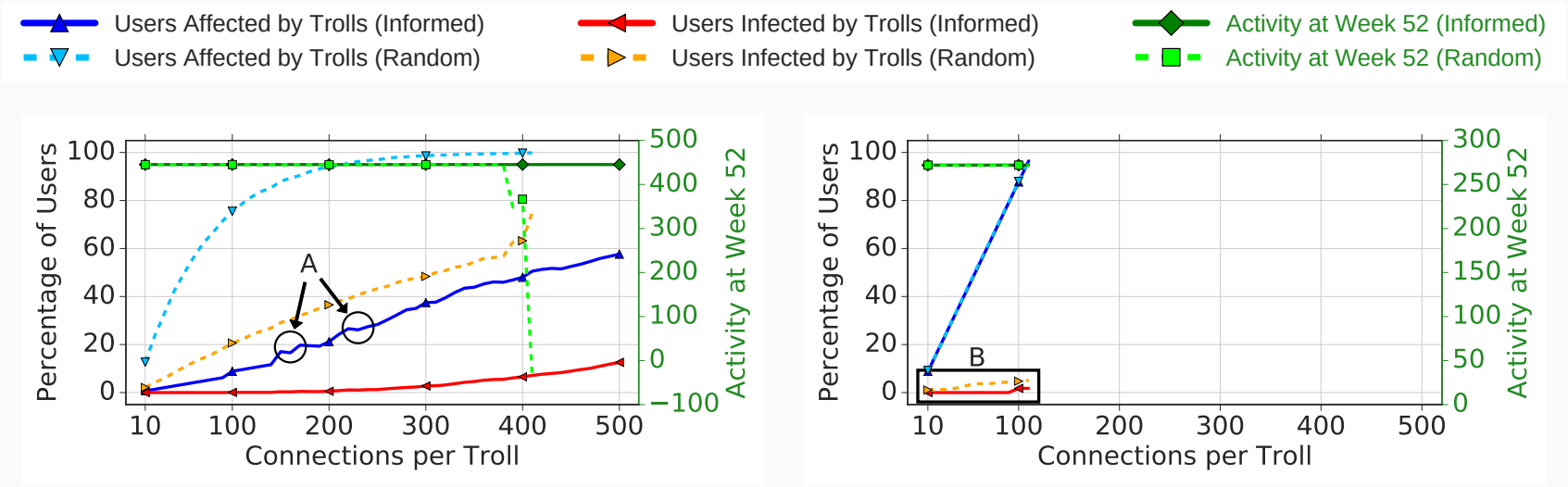


r/Austria



NeuroLex

Increasing Trolls' Exposure



r/Austria

NeuroLex

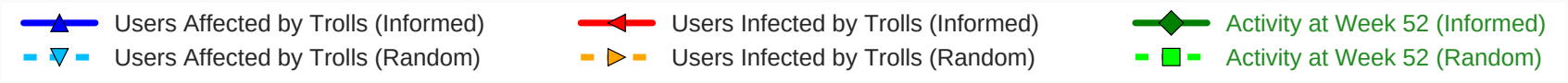
Informed

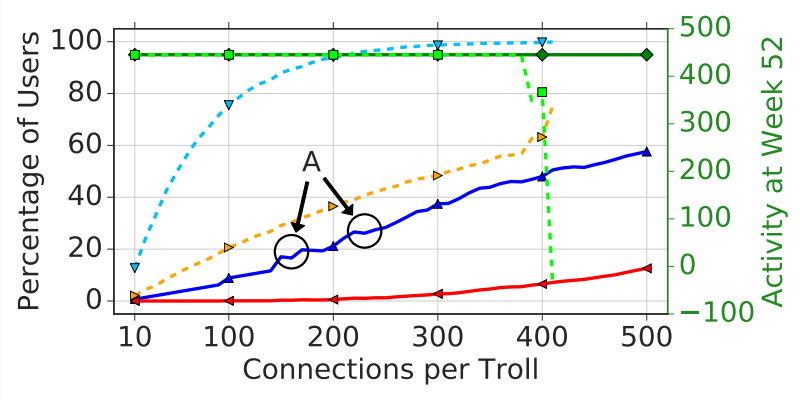
No impact on activity
Temporary decrease of affected users

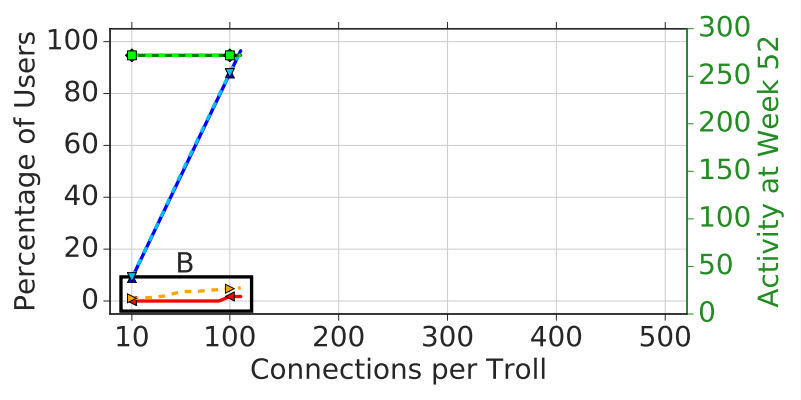
Random

Network collapsed

Increasing Trolls' Exposure







r/Austria

Informed

No impact on activity
Temporary decrease of affected users

Random

Network collapsed

NeuroLex

Informed

No impact on activity
All users affected

Random

No impact on activity
All users affected

Conclusions

Activity Dynamics Model

We showed a novel application of the model and that it is easy to adopt for such tasks.

Adding Trolls

Lower numbers of trolls have higher impact when connecting to the periphery of networks.
High numbers of trolls have higher impact when focusing on central users.

Tipping Point

There seems to be a tipping point at which highly active users cannot longer compensate for the trolls negative influence.

Future Work

Evaluation

Find data sets that actually have information about emerging trolls.

Simulation of Similar Events

For example, joining or leaving users, users that start or end collaboration.

Improvemt of Activity Dynamics Model

Implement filters (e.g.: alpha beta) to increase performance;
Investigate the definition of collaboration networks.

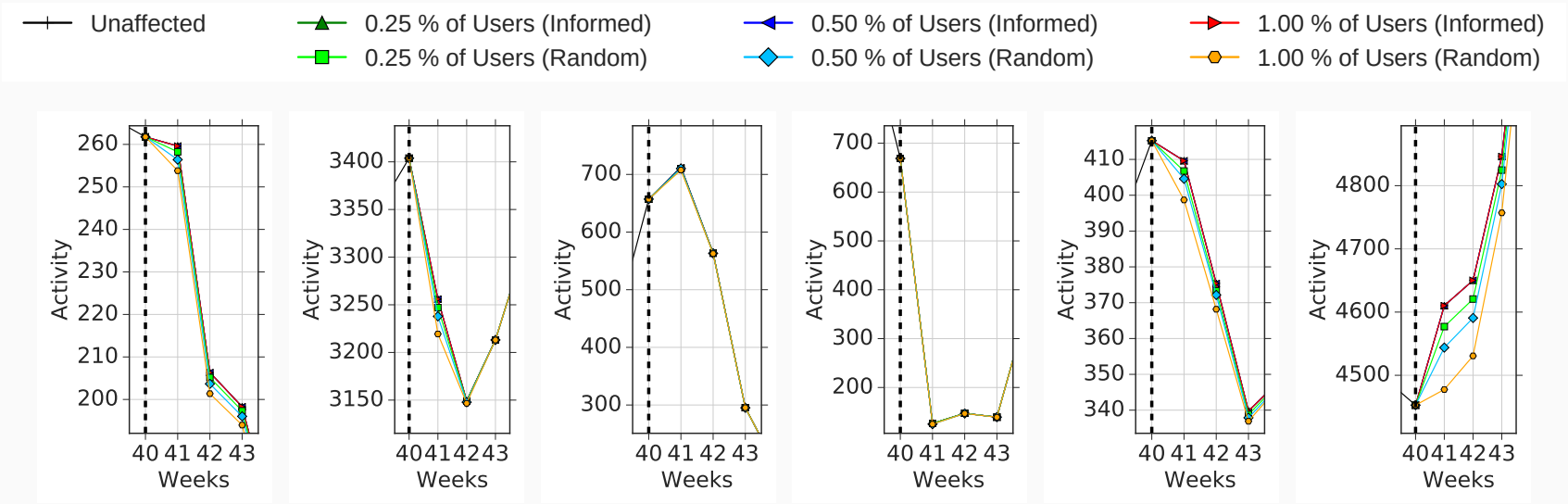
Other Collaboration Network Types

Find other types of collaboration networks.

Questions?

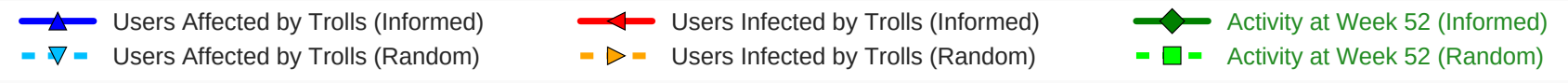
Questions?

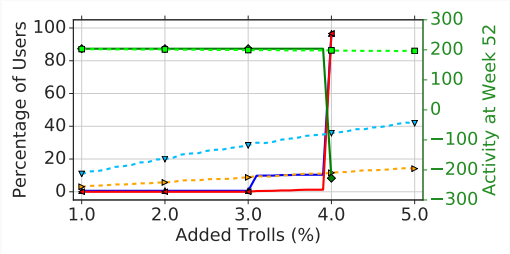
Adding Trolls (Part I)



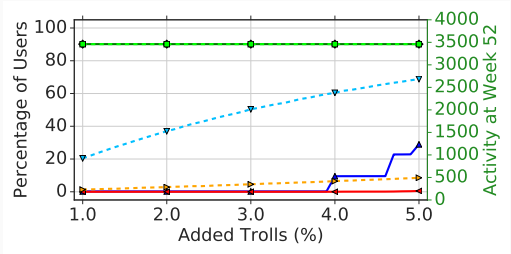
(a) BitcoinStackExchange
(b) EnglishStackExcahnge
(c) DotaWiki
(d) NeuroLex
(e) r/Austria
(f) r/StarWars

Adding Trolls (Part II)

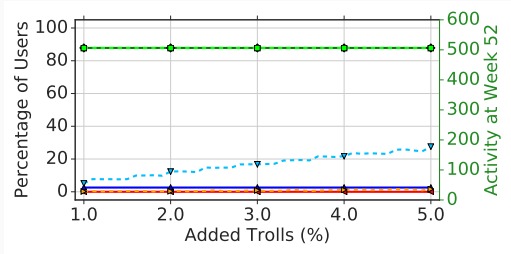




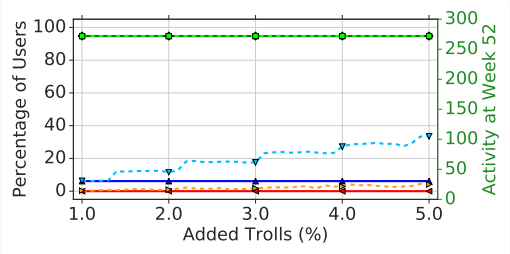
(a) BitcoinStackExchange



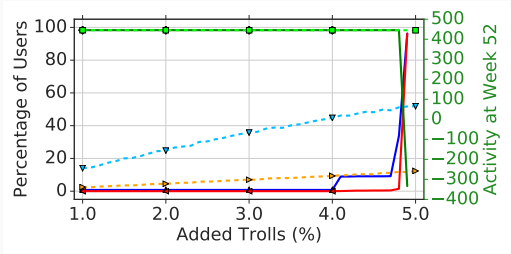
(b) EnglishStackExcahnge



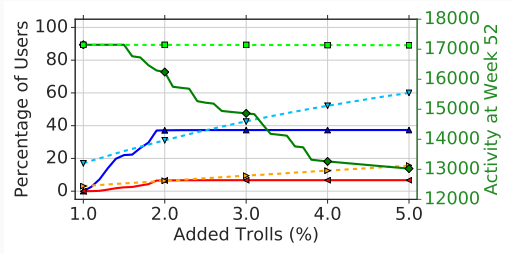
(c) DotaWiki



(d) NeuroLex

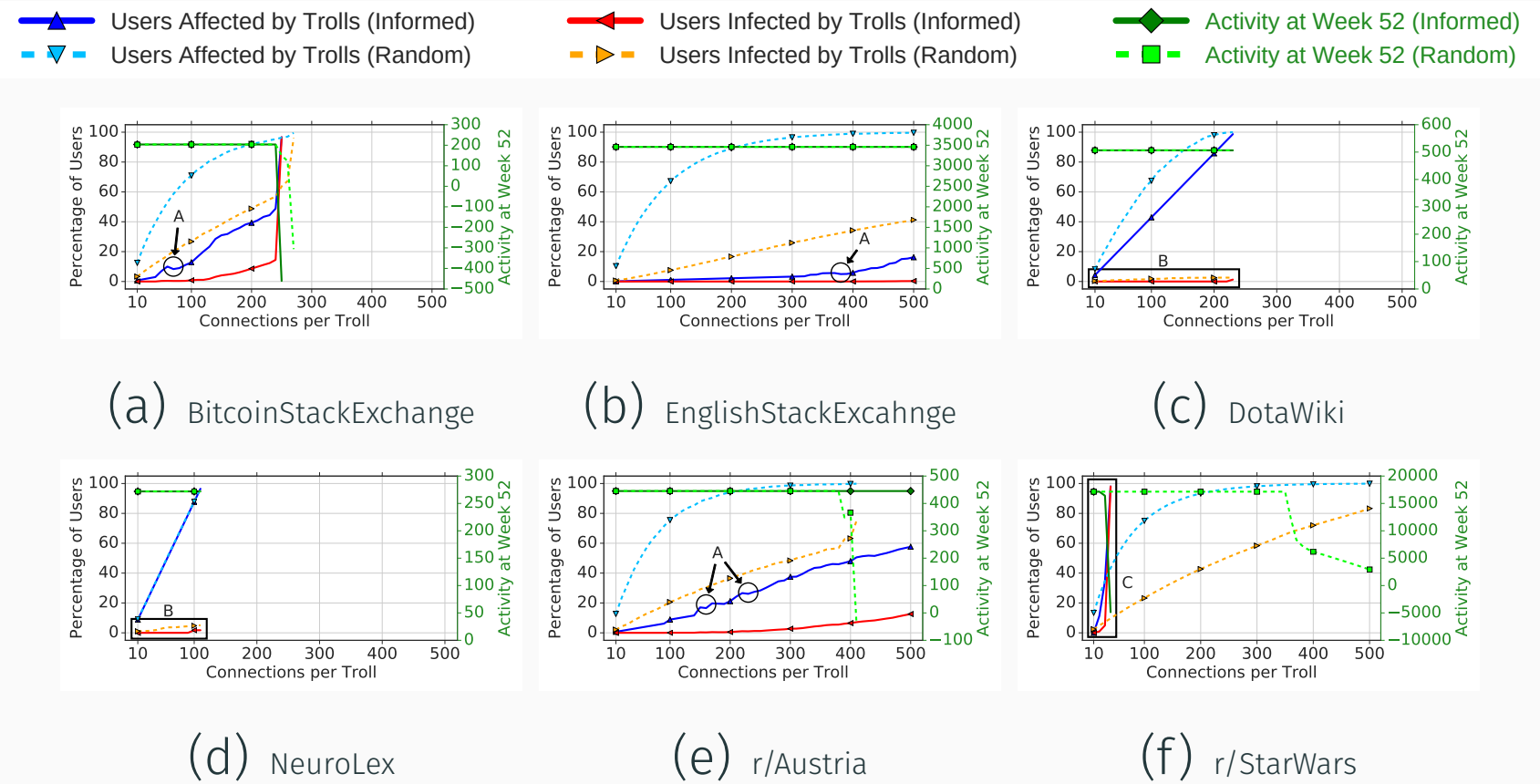


(e) r/Austria

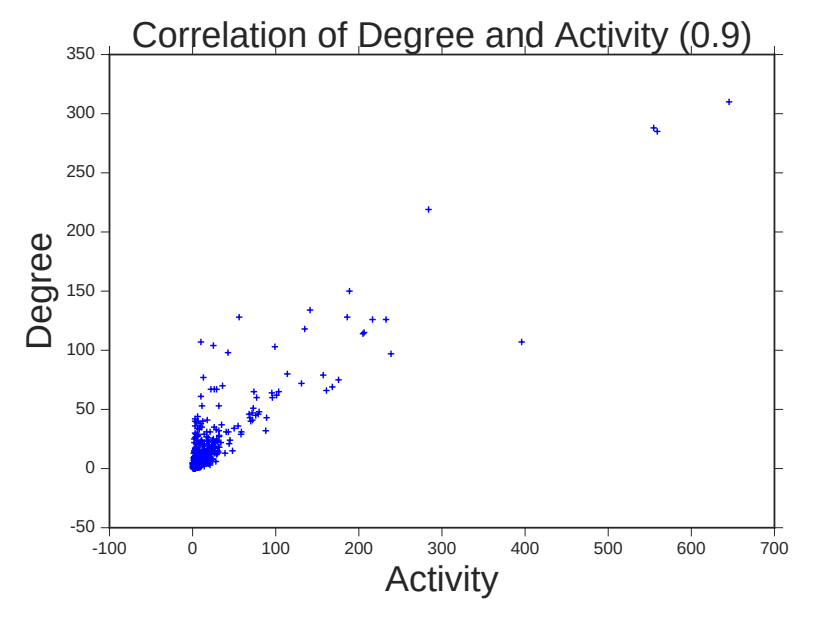


(f) r/StarWars

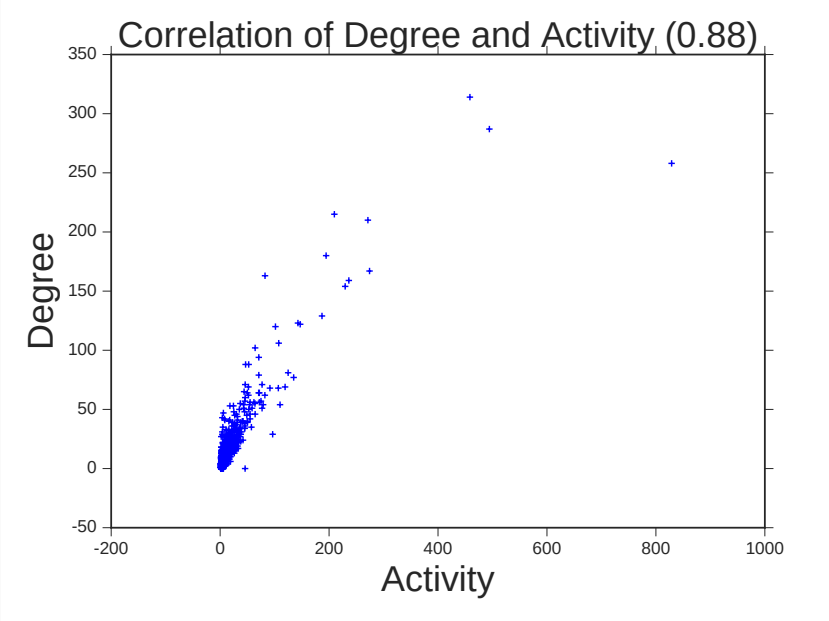
Increasing Trolls' Exposure



Degree/Activity Correlation



(a) BitcoinStackExchange



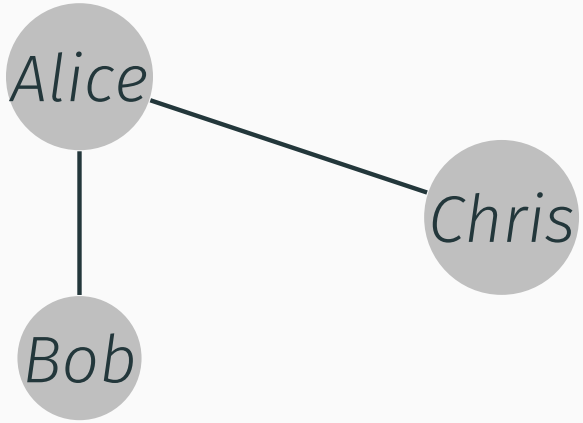
(b) r/Austria

Activity Dynamics Model

Alice collaborates with Bob and Chris.

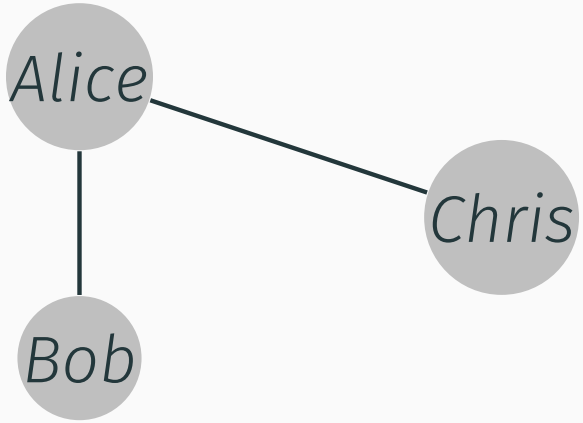
Activity Dynamics Model

Alice collaborates with Bob and Chris.



Activity Dynamics Model

Alice collaborates with Bob and Chris.



$$A = \begin{array}{c} \\ Alice \\ Bob \\ Chris \end{array} \begin{array}{c} \begin{array}{ccc} Alice & Bob & Chris \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$

Activity Dynamics Model

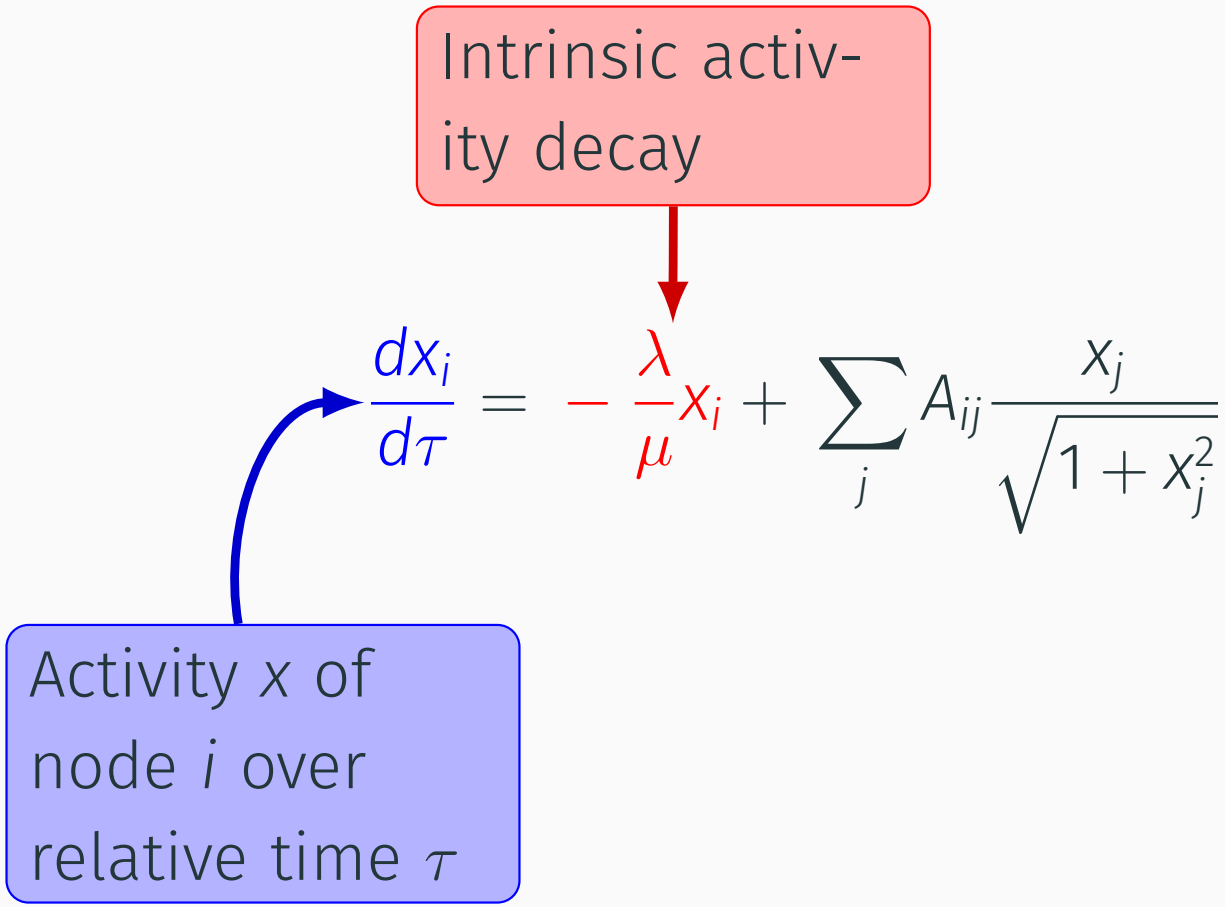
$$\frac{dx_i}{d\tau} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}}$$

Activity Dynamics Model

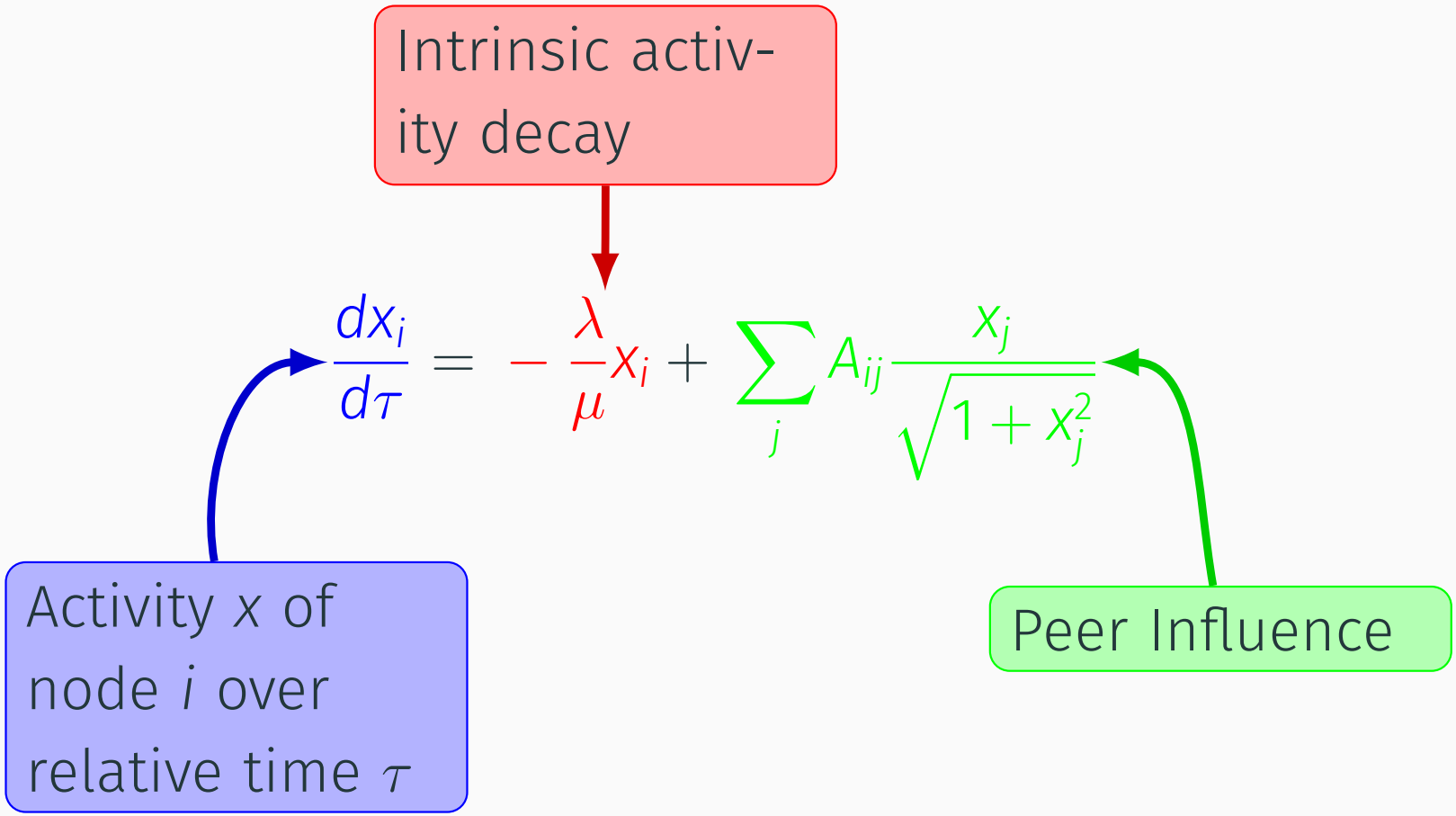
$$\frac{dx_i}{d\tau} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}}$$

Activity x of node i over relative time τ

Activity Dynamics Model



Activity Dynamics Model



User Activity

Dataset	StackExchange		Semantic MediaWikis		Subreddits	
	Bitcoin	English	DotaWiki	NeuroLex	r/Austria	r/StarWars
Mean Activity	11	20	73	319	11	10
Median Activity	3	3	3	7	4	4