Pre-Processing of Dynamic Networks*

Its Impact on Observed Communities and Other Analytics

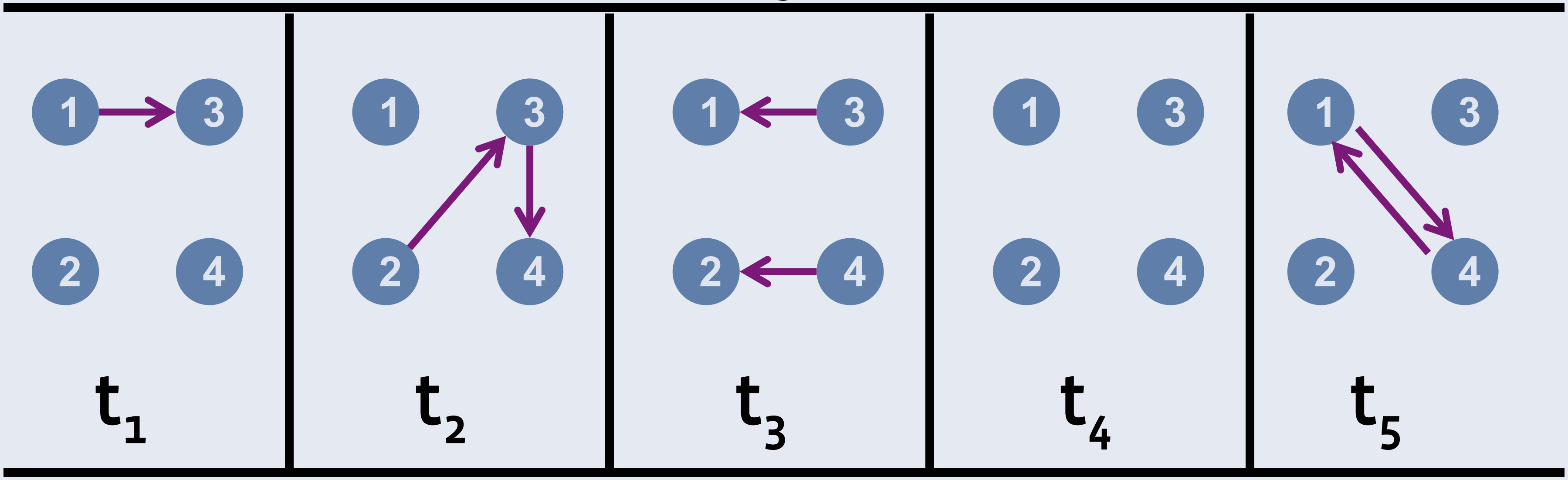
Sofus A. Macskassy

Data Scientist, Facebook

"Unifying Theory and Experiment for Large-Scale Networks"
Simons Institute for the Theory of Computing, UC Berkeley
November 20, 2013

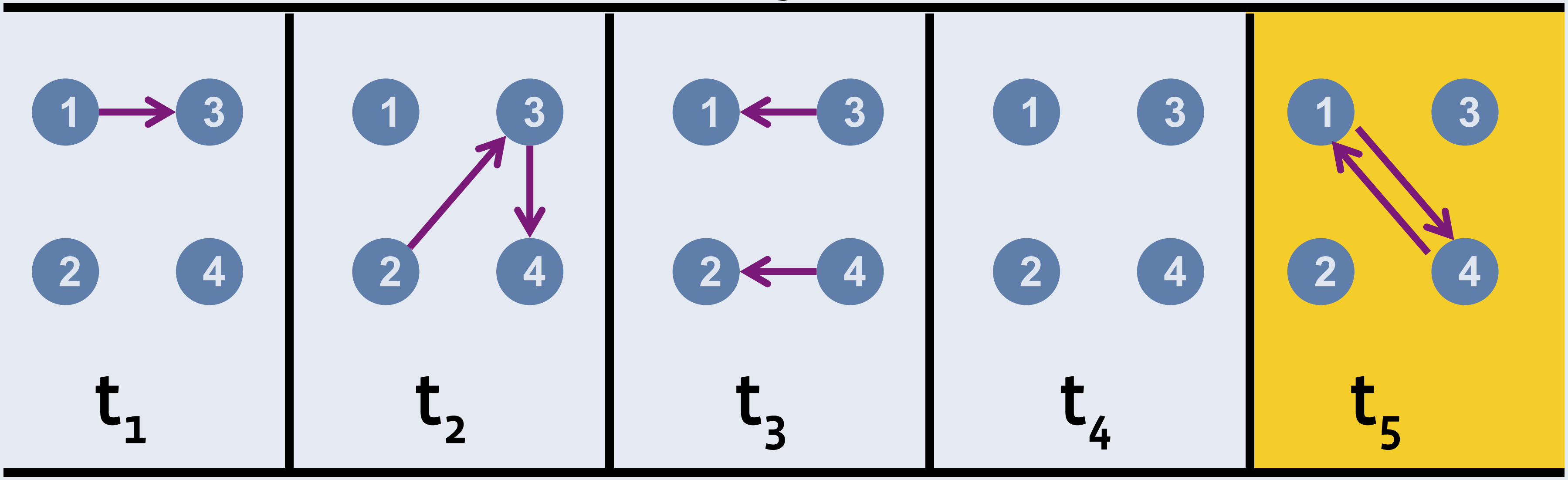
* Work done while at USC/ISI

One problem with dynamic networks

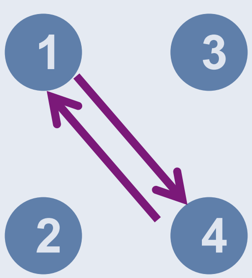


- What does the network look like at time t_5?

One problem with dynamic networks

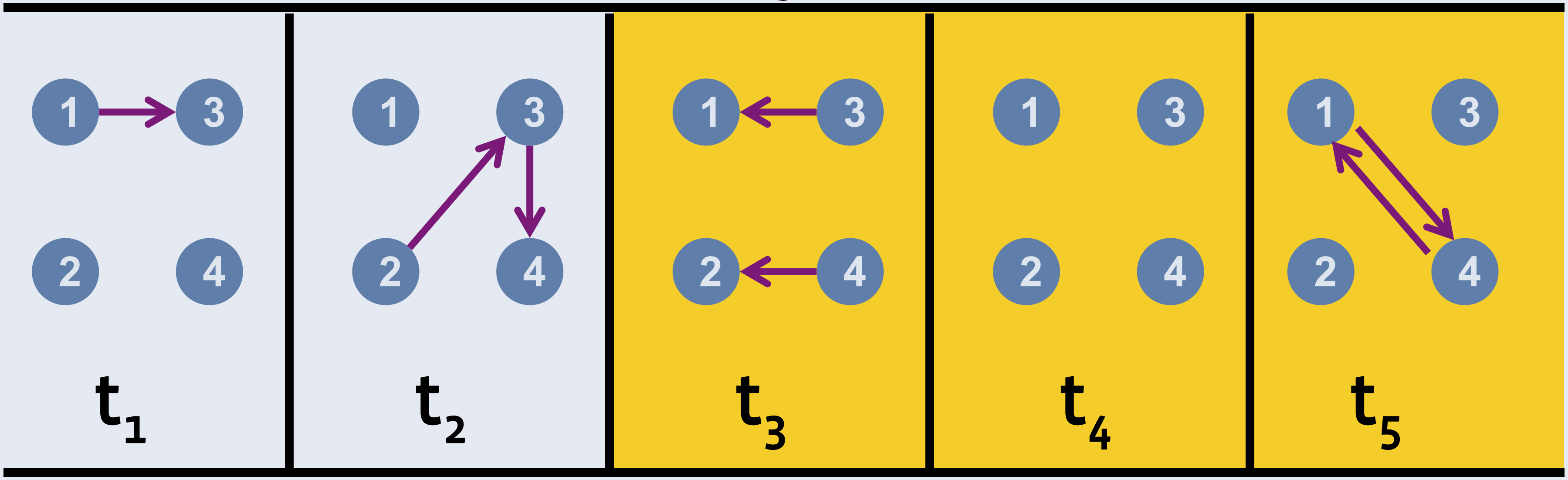


- What does the network look like at time t_5?

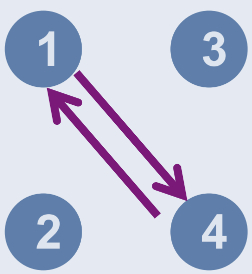


Latest snapshot?

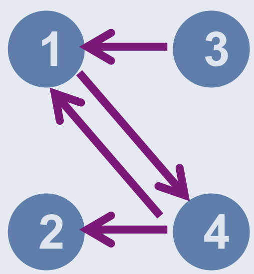
One problem with dynamic networks



- What does the network look like at time t_5?

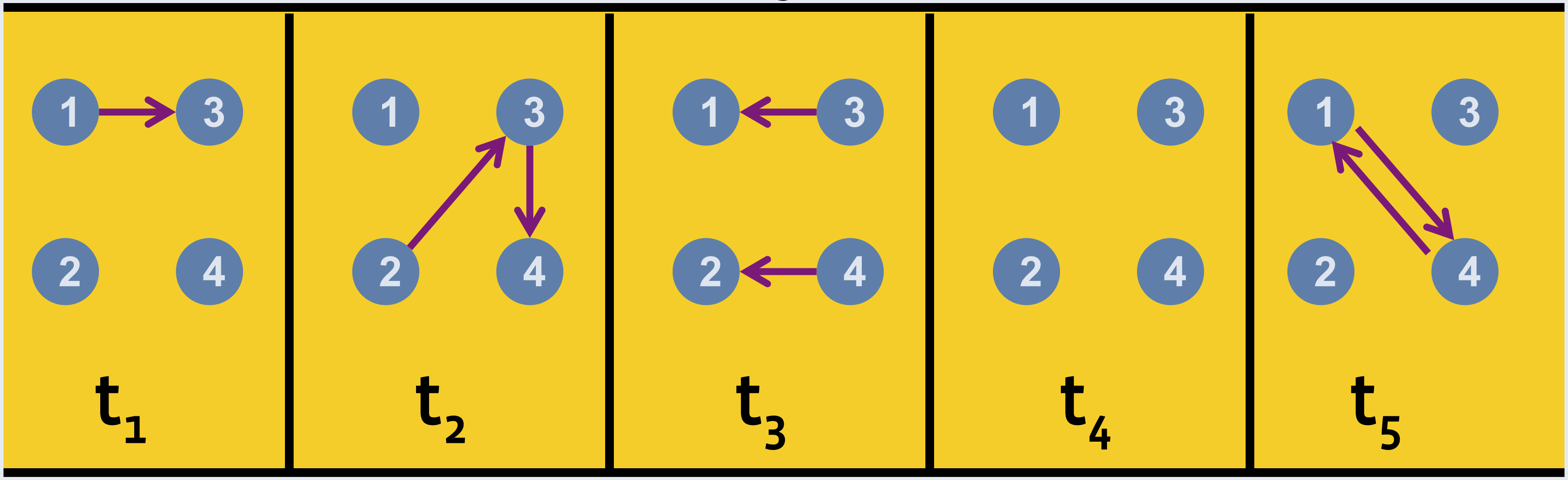


Latest snapshot?

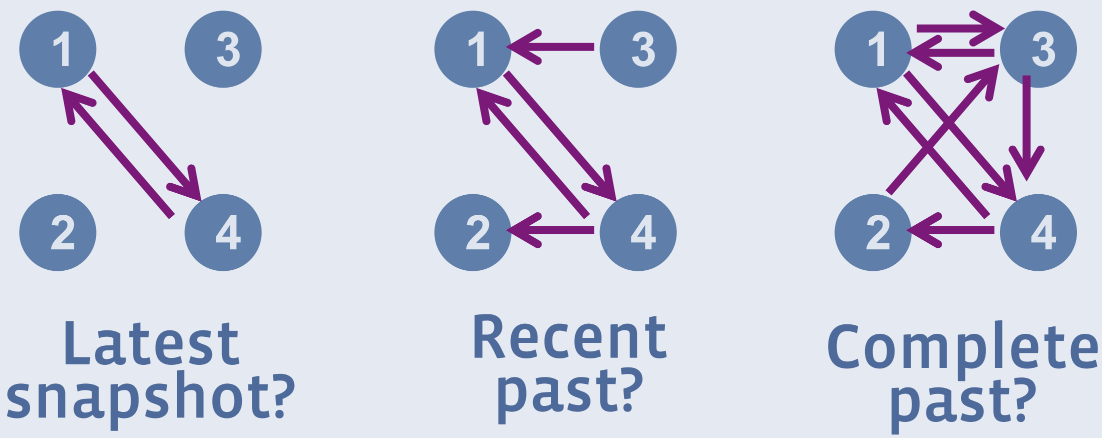


Recent past?

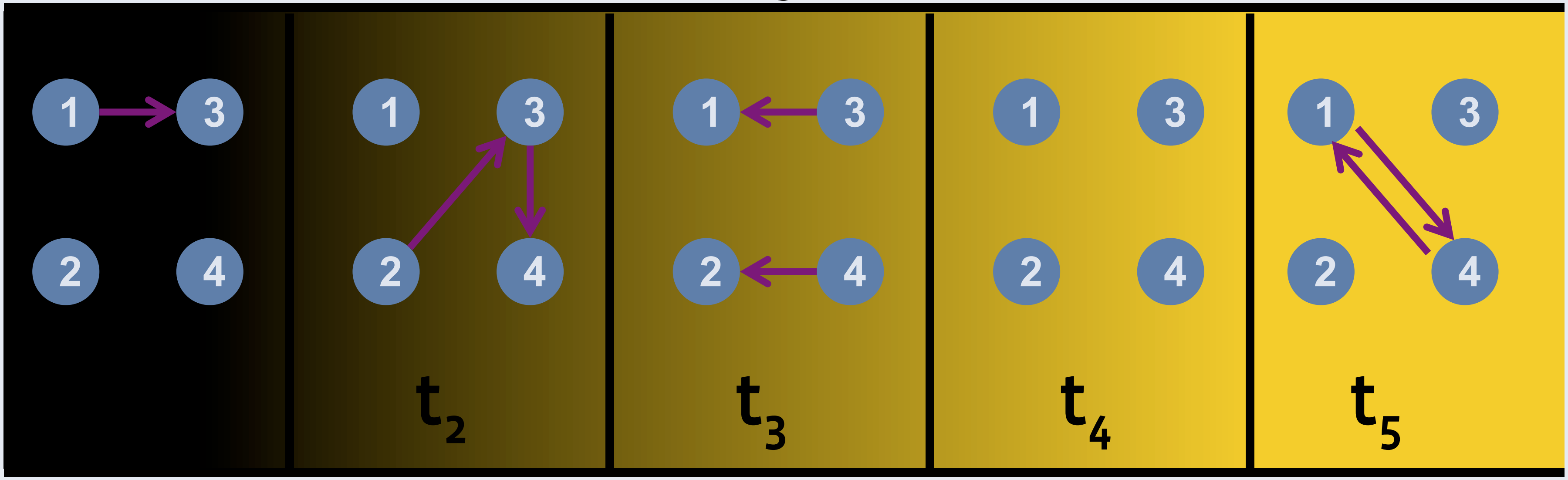
One problem with dynamic networks



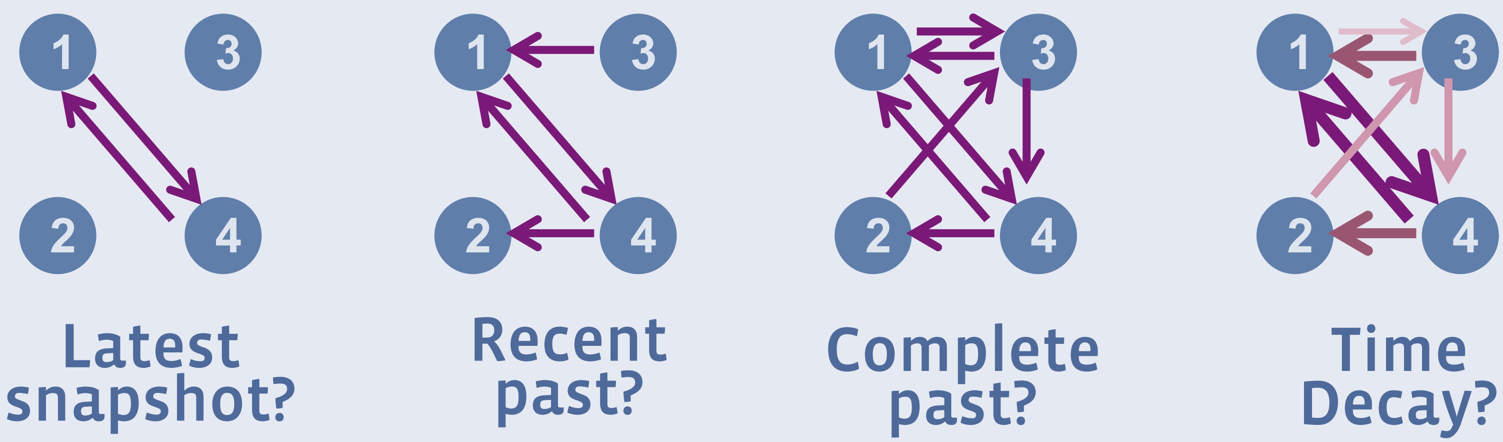
- What does the network look like at time t_5?



One problem with dynamic networks

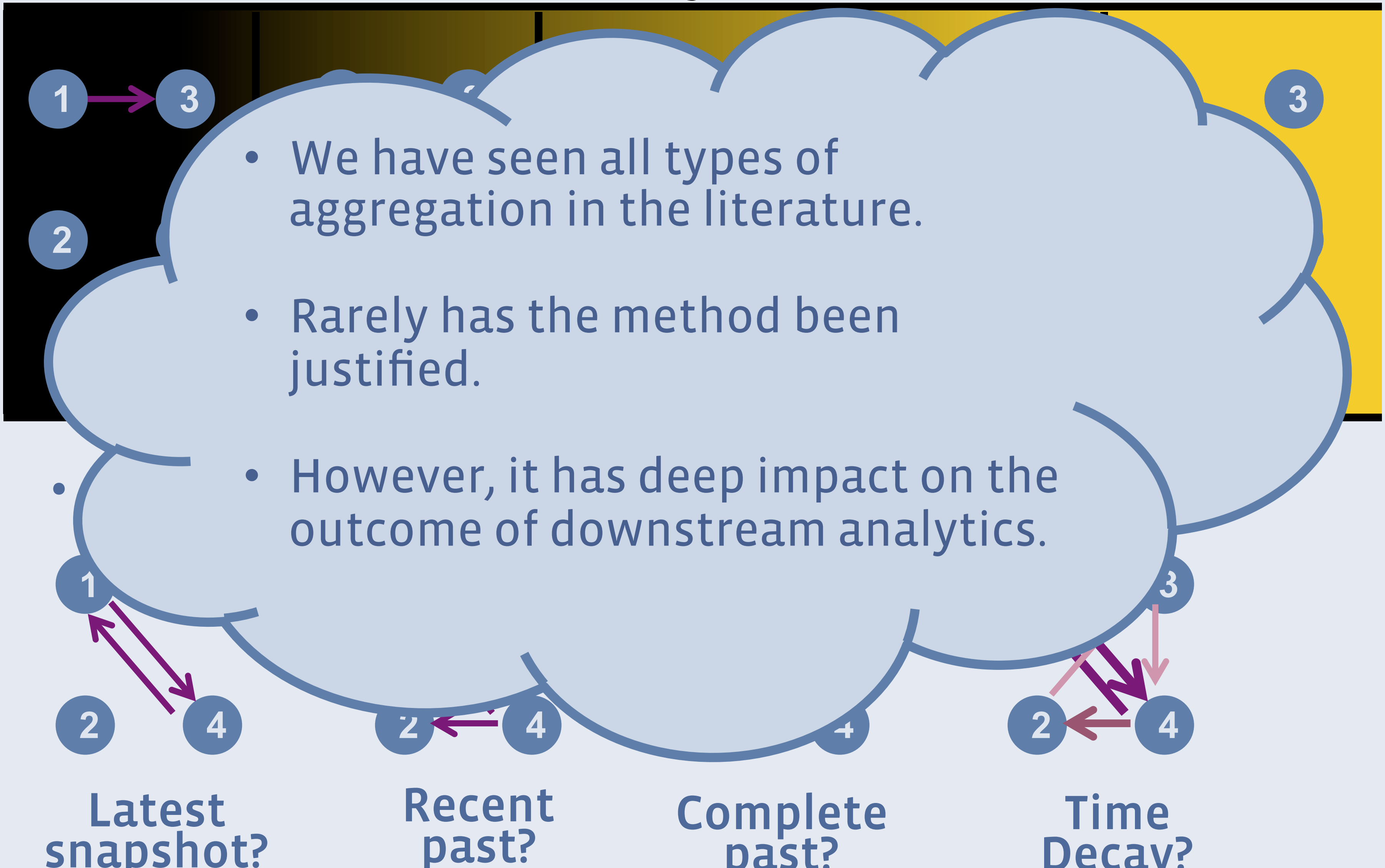


- What does the network look like at time t_5?



One problem with dynamic networks

- We have seen all types of aggregation in the literature.
- Rarely has the method been justified.
- However, it has deep impact on the outcome of downstream analytics.



How does aggregation affect analysis?

We here explore four particular questions:

1. What does the network look like at time t?
2. What are the communities and how do they evolve?
3. How do nodes change (centrality/membership)?
4. What is the impact on analytics?

   (e.g., on analytics such as machine learning)

Generating Network at Time *t*

What is a "current" network?

- At any given $t \pm \varepsilon$, may be few, if any, edges
- Solution: aggregate over a time window δ

However, past edges are also informative

- Edges from prior window may still have some influence
- Add edges from prior network with decay parameter α
- Prune edges with low weight (below η)

Generating Network at Time *t*

Network at time *t*, then can be defined as

- Adjacency matrix at time ***t***:

$$\mathbf{A}_0^t = \left\{ e_{ij}^{t'} \middle| (t-\delta) \le t' \le t \right\}$$

$$\mathbf{A}^t = \mathbf{A}_0^t + \alpha \cdot \mathbf{A}^{(t-\delta)}$$

- Final network at time ***t***:

$$\mathbf{G}^t = \left(\mathbf{V}^t, \mathbf{E}^t \right)$$

$$\mathbf{E}^t = \left\{ e_{ij}^t \middle| a_{ij}^t \ge \eta \right\}, a_{ij}^t \in \mathbf{A}^t$$

$$\mathbf{V}^t = \left\{ v_i \middle| \exists j \left(e_{ij} \in \mathbf{E}^t \wedge e_{ji} \in \mathbf{E}^t \right) \right\}$$

- Normalize ***t*** and δ to result in snapshots $\mathbf{G}^1 \ldots \mathbf{G}^T$

Tracking Communities

Given $\mathbf{G}^t$, use *modularity clustering* to identify *k* communities (using weighted or unweighted edges)

$$\mathbf{C}^t = \left(c_1^t, \cdots, c_k^t\right), c_i^t = \left\{v_j \middle| v_j \in \mathbf{V}^t\right\}$$

Identify communities from $\mathbf{C}^{t-1}$ which are also in $\mathbf{C}^t$

Categorize community actions into four major events

Continue: $\left|c_i^{(t-1)} \cap c_j^t\right| > 0.5 * \left|c_i^{(t-1)}\right|$ and $\left|c_i^{(t-1)} \cap c_j^t\right| \geq 0.65 * \left|c_j^t\right|$

Merge: $\left|c_i^{(t-1)} \cap c_j^t\right| > 0.5 * \left|c_i^{(t-1)}\right|$ and $\left|c_i^{(t-1)} \cap c_j^t\right| < 0.65 * \left|c_j^t\right|$

Split: Significant portions (>30%) of $c_i^{(t-1)}$ move into two or more communities in $\mathbf{C}^t$

Death: None of the above

Tracking Nodes

- From community actions, we can track nodes
- Split node movement into three major events

Stay: Community continues/merges and node stays with community

Leave: Community continues/merges but node goes to another community

Other: Community splits or dies

Experimental Study

- Research question:

  What is the effect of varying graph-extraction parameters?

- Methodology:

  1. Select data sets
  2. Vary parameters and extract communities
  3. Track communities over time
  4. Downstream analytics: machine learning

Data Sets used

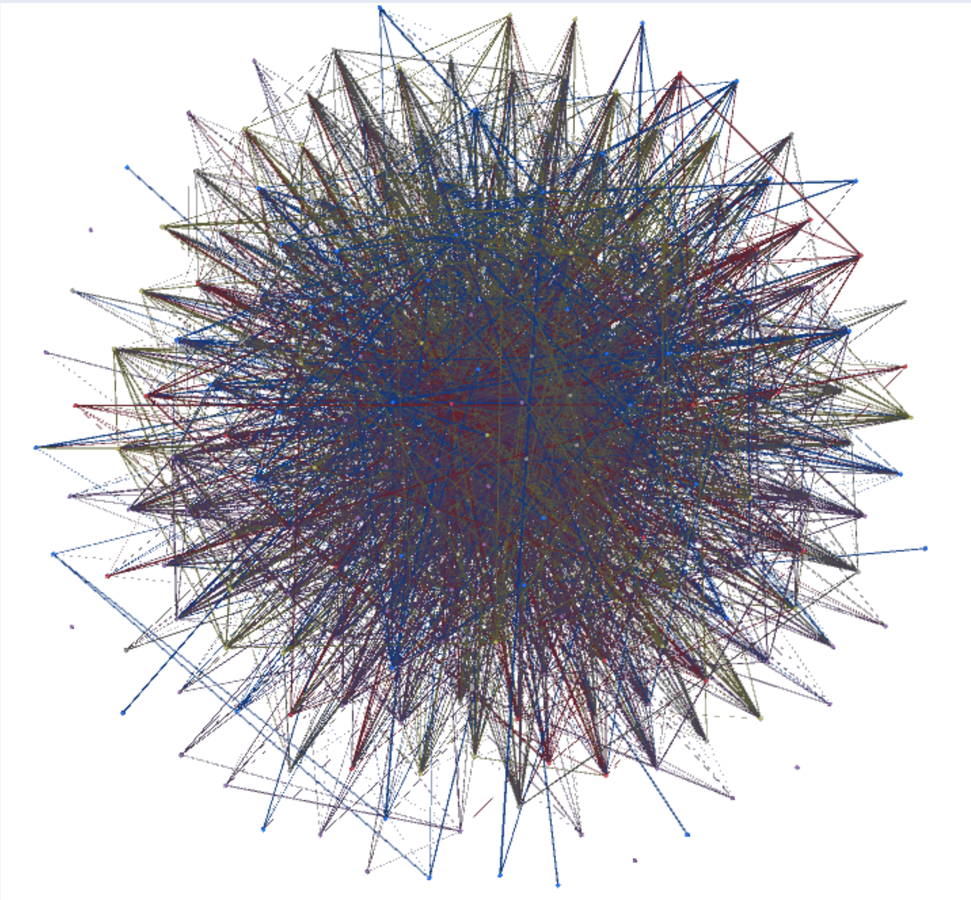
- **The Enron email data set** **(http://www.cs.cmu.edu/~enron/)**
  - 151 nodes (Enron employees) communicating with each other over 2 years
  - We set **δ =1 month** for our study
  - Edge-weight = # emails between people in a given month
  - 60K emails, containing 139K links over 2 years
- **World trade flows (WTF)** **(http://www.nber.org/data/)**
  - 203 nodes (countries) of trades between countries from 1962 through 2000
  - We set **δ =1 year** as that is the granularity of the data
  - Edge-weight of $i \rightarrow j$ is normalized across all $i \rightarrow k$ (keep top-10)
  - 32K links

Varying parameters

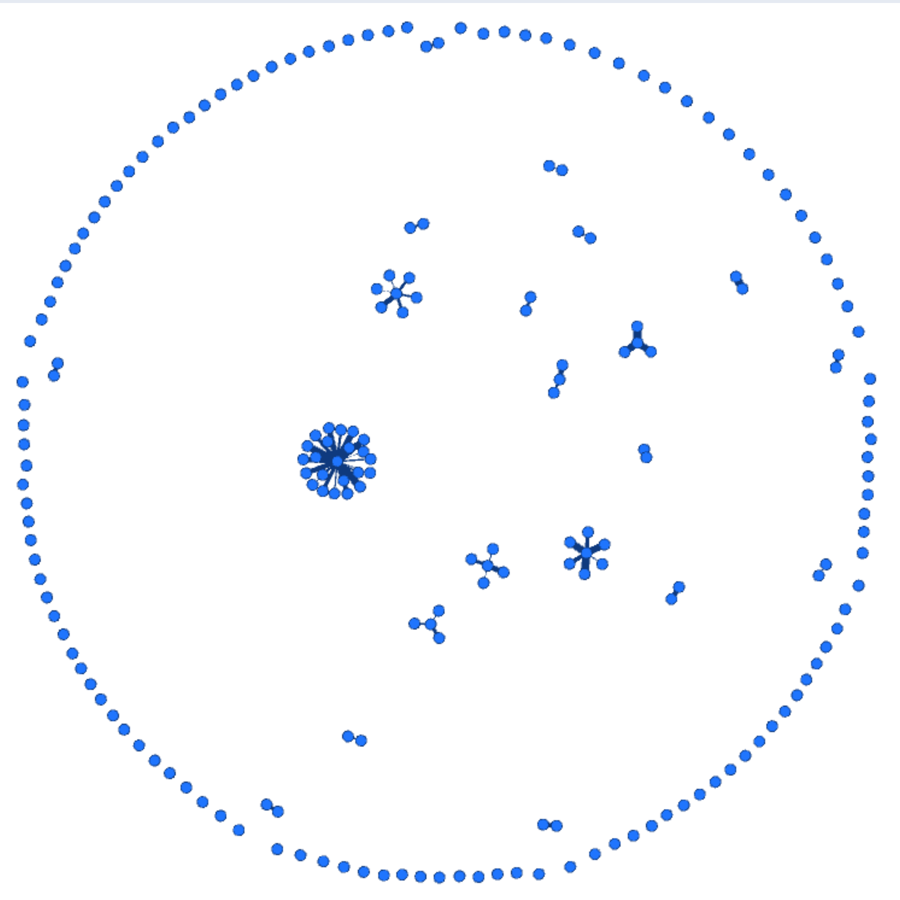
- We performed an empirical study looking at the effects of changing α and η
  - α ={0.5, 0.75, 0.9, 1.0}
  - η ={0.05, 5.0, 10.0} [enron]
  - η ={0.01, 0.05, 0.10, 0.25, 0.5, 0.75, 1.0} [world trade flows]
  - We tested when using weighted and unweighted edges
  - Used to generate attribute values and identifying communities

What does the network look like? World trade flows (1993)

(α=1.0; η=0.05)

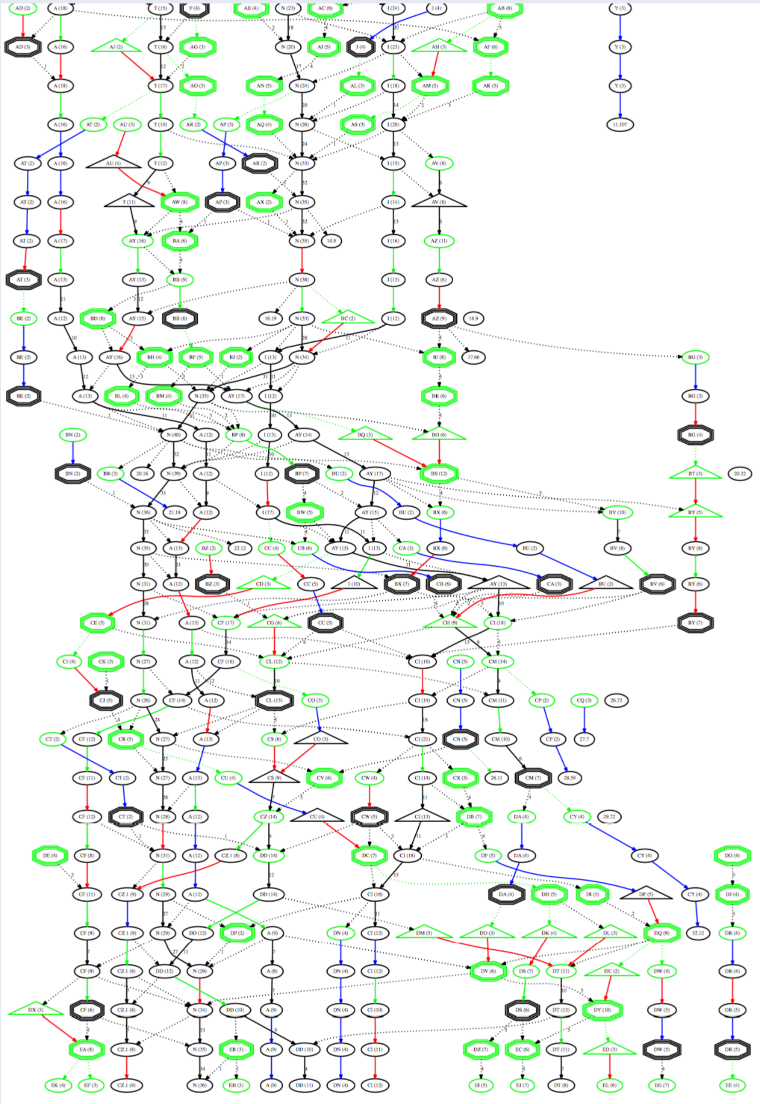


(α=0.5; η=0.75)

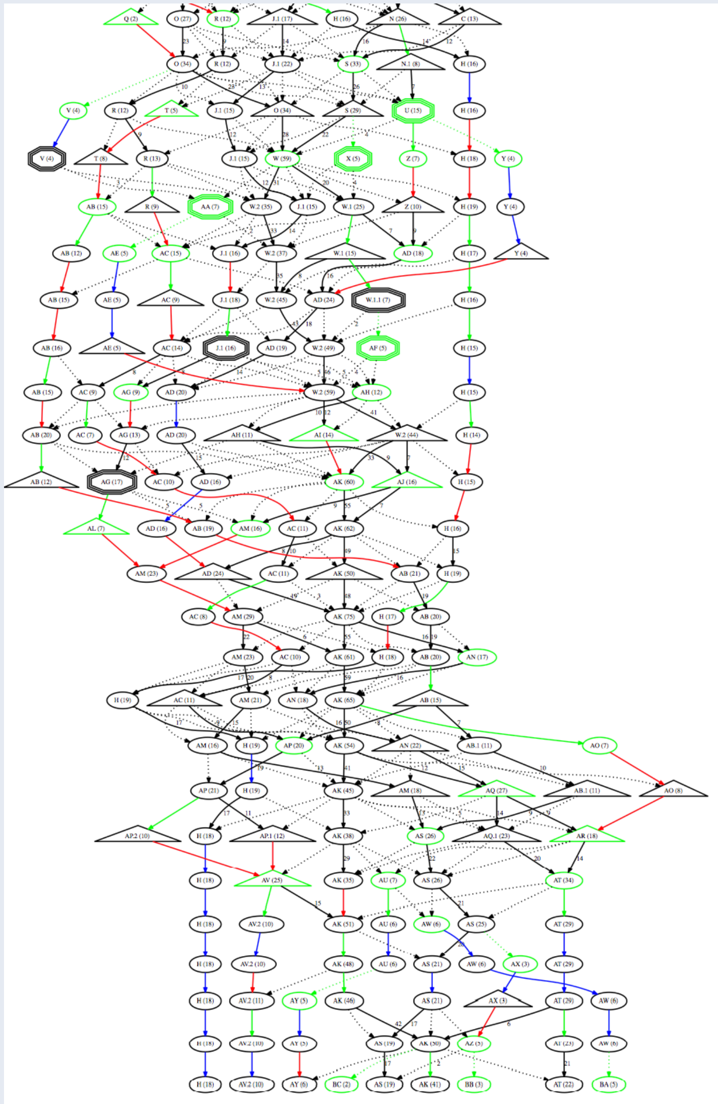


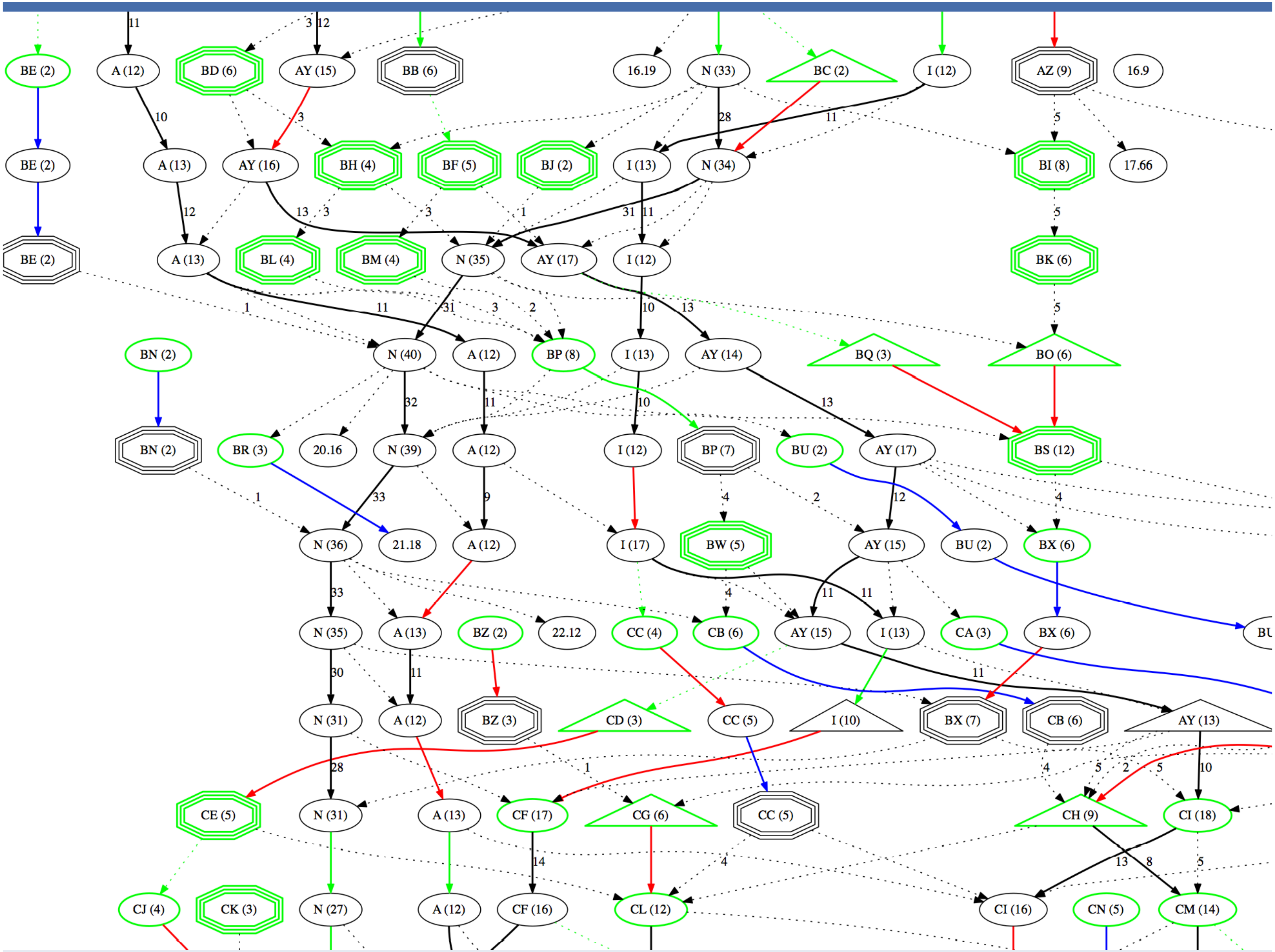
Snapshot of evolution… (α=0.5; η=0.05)

World Trade Flows



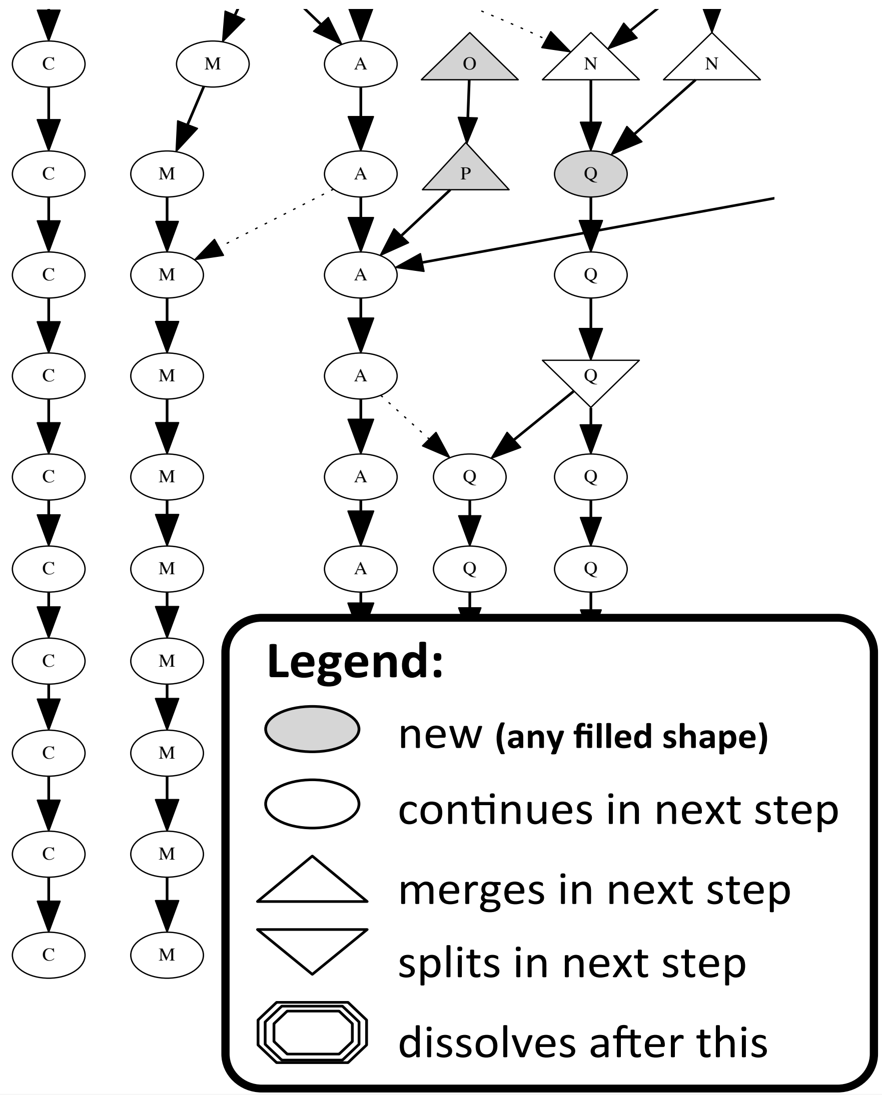
Enron




BE (2)
A (12)
BD (6)
AY (15)
BB (6)
16.19
N (33)
BC (2)
I (12)
AZ (9)
16.9
BE (2)
A (13)
AY (16)
BH (4)
BF (5)
BJ (2)
I (13)
N (34)
BI (8)
17.66
BE (2)
A (13)
BL (4)
BM (4)
N (35)
AY (17)
I (12)
BK (6)
BN (2)
N (40)
A (12)
BP (8)
I (13)
AY (14)
BQ (3)
BO (6)
BN (2)
BR (3)
20.16
N (39)
A (12)
I (12)
BP (7)
BU (2)
AY (17)
BS (12)
N (36)
21.18
A (12)
I (17)
BW (5)
AY (15)
BU (2)
BX (6)
N (35)
A (13)
BZ (2)
22.12
CC (4)
CB (6)
AY (15)
I (13)
CA (3)
BX (6)
N (31)
A (12)
BZ (3)
CD (3)
CC (5)
I (10)
BX (7)
CB (6)
AY (13)
CE (5)
N (31)
A (13)
CF (17)
CG (6)
CC (5)
CH (9)
CI (18)
CJ (4)
CK (3)
N (27)
A (12)
CF (16)
CL (12)
CI (16)
CN (5)
CM (14)

Effect on communities stability

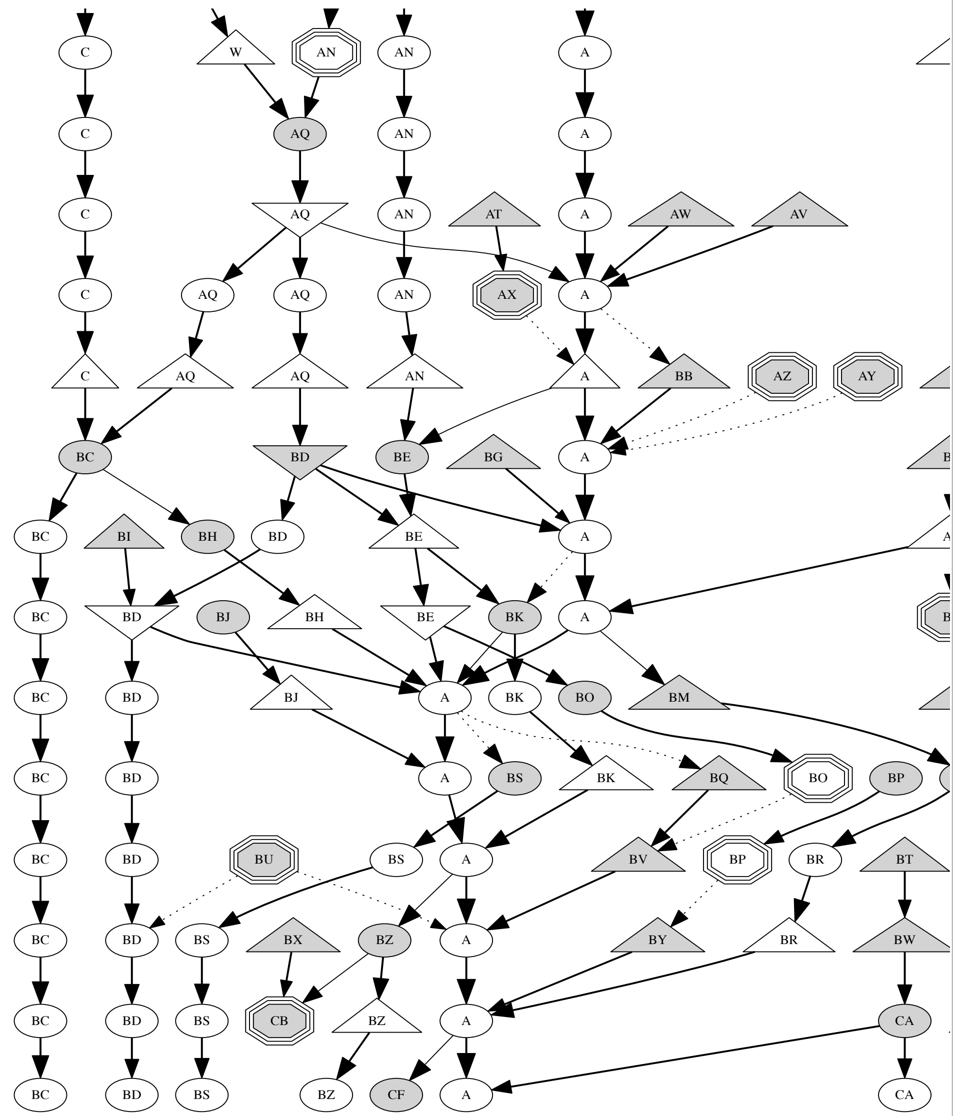
WTF: α=1.0; η=0.05; weighted edges



Legend:

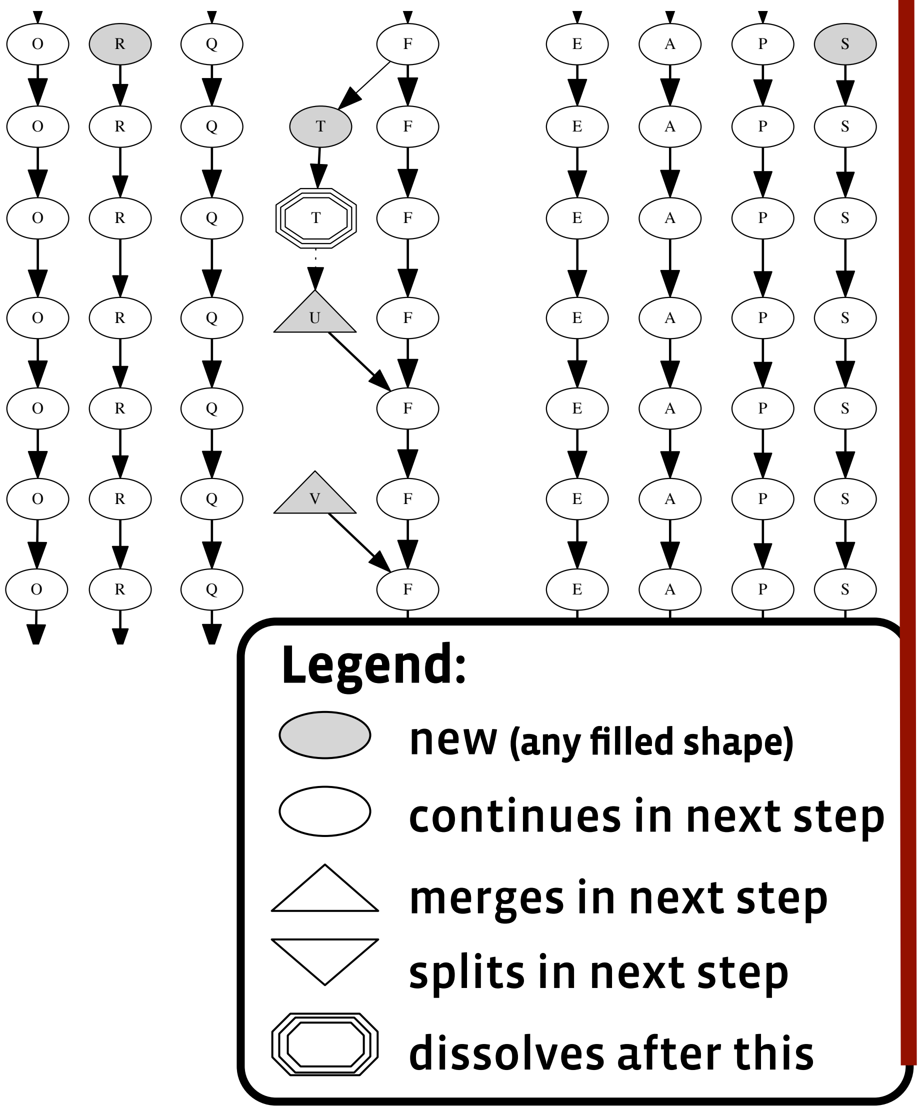
- new **(any filled shape)**
- continues in next step
- merges in next step
- splits in next step
- dissolves after this

WTF: α=0.5; η=0.05; weighted edges

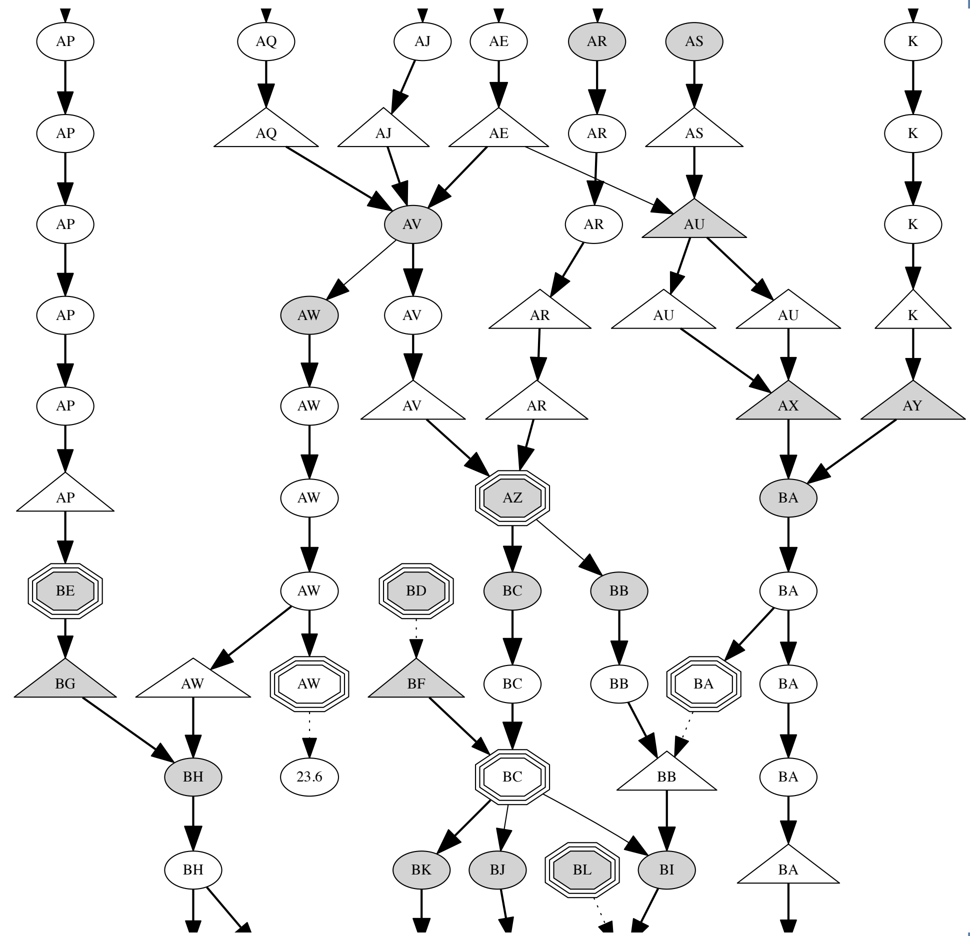


Effect on communities stability

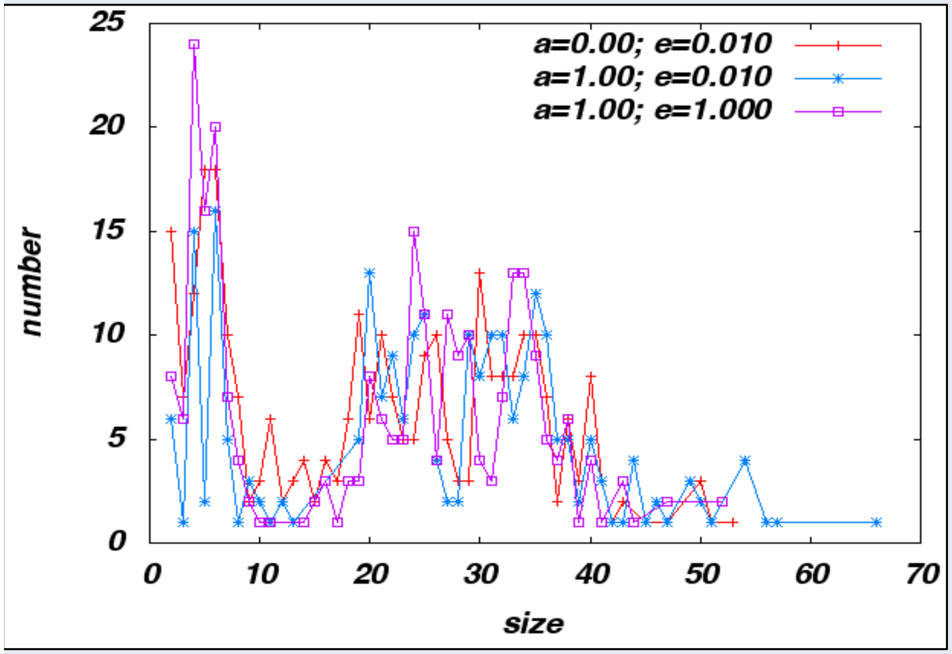
Enron: α=1.0; η=5.0; weighted edges

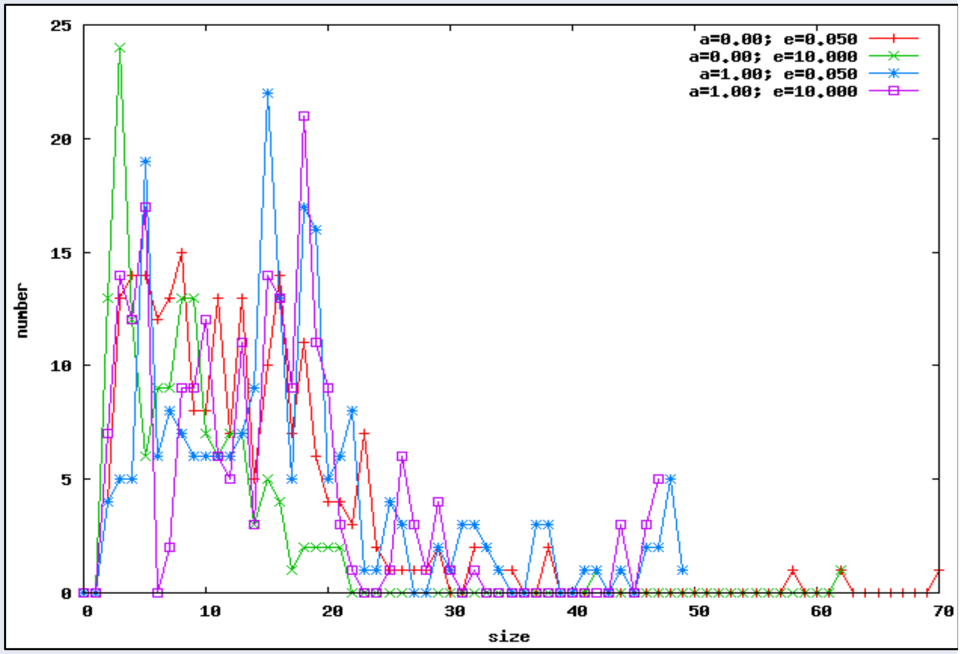


Enron: α=0.5; η=5.0; weighted edges



Effect on community sizes





World Trade Flow

Enron

Effect on longevity of communities

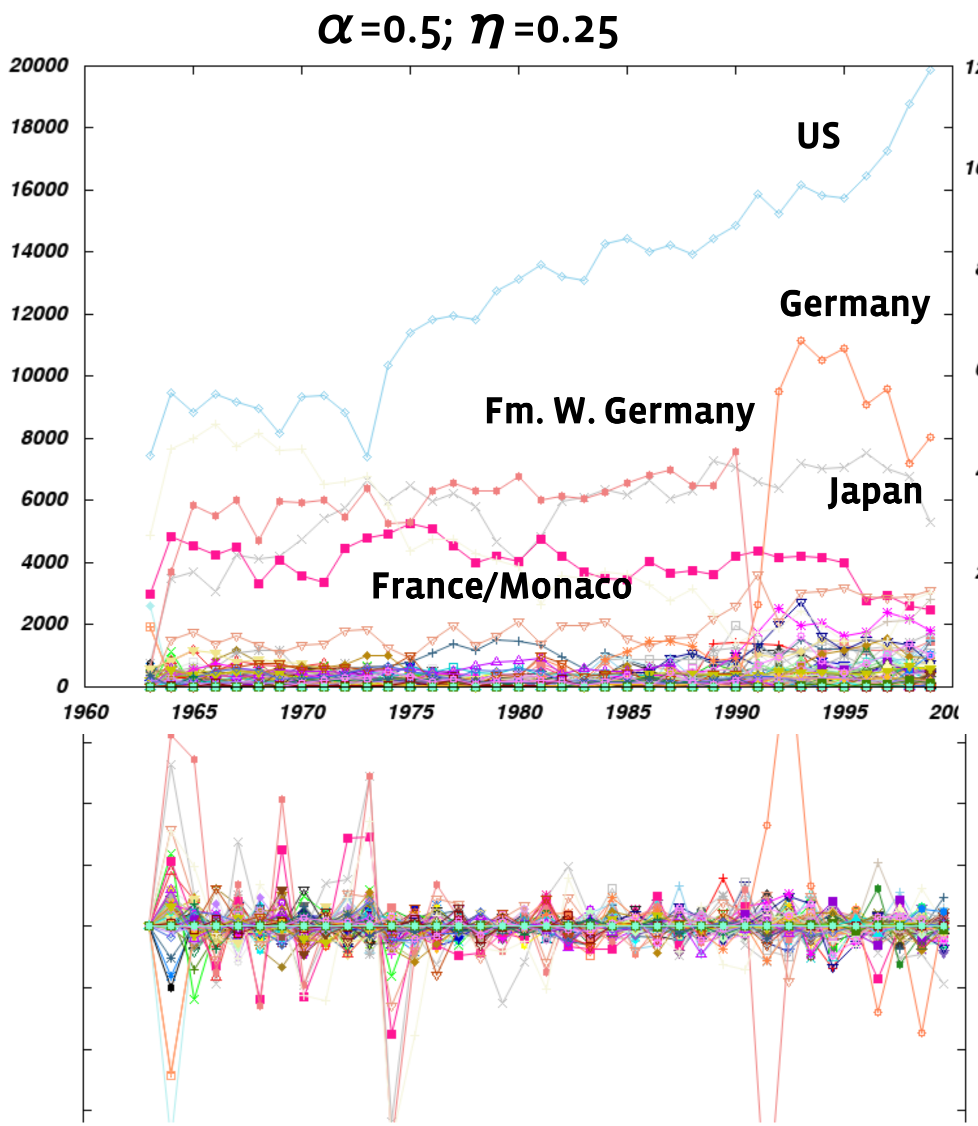
WTF: α=1.0; η=0.05

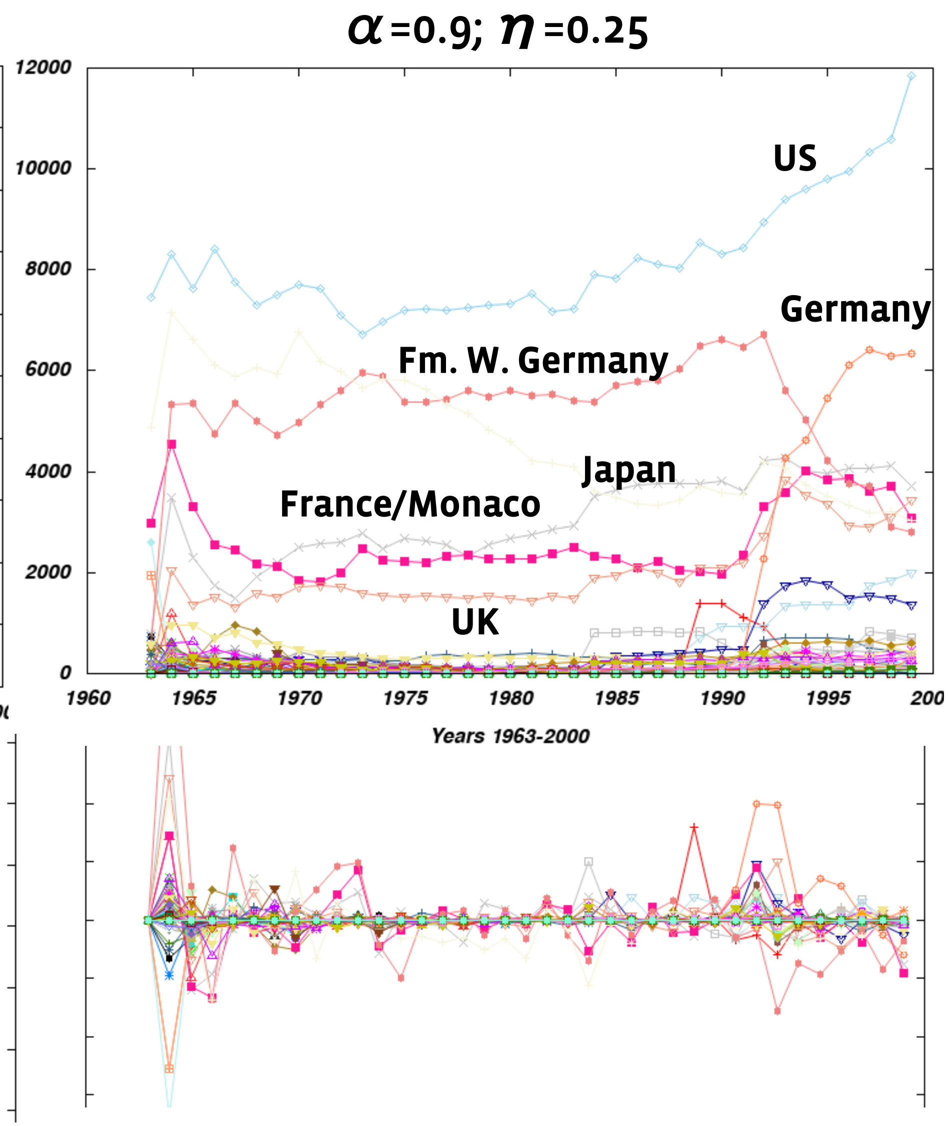
Group 1	Group 2	Group 3	Group 4
38	38	38	37
UK	Canada	Japan	Syria
Ireland	USA	Asian NES	Italy
Cyprus	Colombia	China HK	Jordan
Mauritius	Ecuador	Lao	Austria
Malta	Mexico	Malaysia	Iraq
Bermuda	Costa Rica	Singapore	Czechoslovak
Fiji	El Salvador	Thailand	Germany
Samoa	Guatemala	China	Bulgaria
New Zealand	Honduras	Korea	Hungary
Kenya	Dominican R	Vietnam	Fm. Yugoslav
	Haiti	US NES	Turkey
	Trinidad	Myanmar	Lebanon
	Jamaica	Cambodia	Saudi Arabia
	Peru	Indonesia	Albania
	Venezuela	Philippines	Romania
	Bahamas	Taiwan	Somalia
	St. Pierre		Poland
	Fr. Guiana		
	Guyana		
	Suriname		

WTF: α=0.5; η=0.05

Group 1	Group 2	Group 3
38	31	22
Canada	Japan	UK
USA	Asian NES	Ireland
Colombia	China HK	Cyprus
Ecuador	Lao	Mauritius
Mexico	Malaysia	
Costa Rica	Singapore	
El Salvador	Thailand	Malawi (19)
Guatemala	China	Fiji (18)
Honduras	Korea	Kenya (18)
Dominican R	Vietnam	
Haiti		
Trinidad		
	US NES (27)	
	Myanmar (25)	
Jamaica (36)	Kiribati (26)	
Venezuela (34)	Papua N. Guin (26)	
Bahamas (34)		
Peru (33)		

Effect on Betweenness centrality (world trade flow)





Downstream Analytics: Machine Learning

- 2 Classification problems
  - Given $G^1 \ldots G^{(t-1)}$ predict changes in communities going into C^t
  - Given $G^1 \ldots G^{(t-1)}$ predict changes in nodes going into C^t
- Attributes used are purely
  - Community: Density, inter- and intra-link ratio, size, number of triangles, average closeness centrality, …
  - Nodes: Number of triangles, inter- vs intra-link ratio, size of communities linked to, …
- We performed 5x2 CV
- Various off-the-shelf ML methods used
  - Logistic regression, decision trees, naïve bayes

Class Distribution: Enron

Enron communities

	C / M / S	C / M / S	C / M / S
α \ η	0.05	5.00	10.00
0.50	151 / 52 / 2	124 / 55 / 3	110 / 56 / 2
0.75	176 / 20 / 1	179 / 34 / 3	180 / 32 / 2
0.90	180 / 22 / 1	187 / 25 / 1	196 / 25 / 1
1.00	200 / 16 / 0	197 / 16 / 1	204 / 17 / 1

Enron nodes

	L / S	L / S	L / S
α \ η	0.05	5.00	10.00
0.50	737 / 3259	660 / 1908	485 / 1365
0.75	505 / 3633	545 / 2926	463 / 2462
0.90	392 / 3741	406 / 3443	331 / 3128
1.00	320 / 3822	326 / 3612	266 / 3449

Class Distribution: WTF

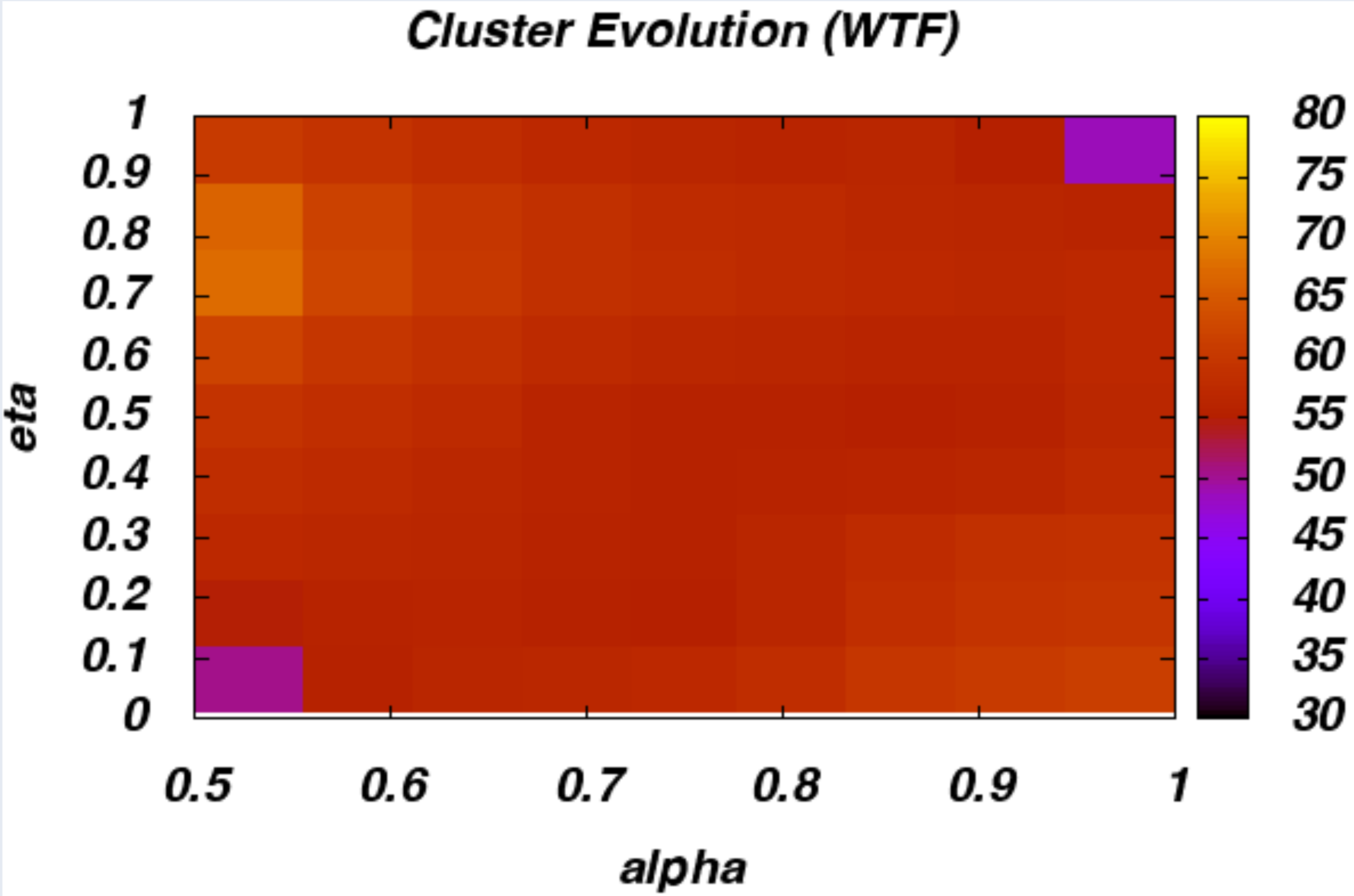
World trade flows communities

	C / M / S	C / M / S	C / M / S	C / M / S	C / M / S	C / M / S	C / M / S
α \ η	0.01	0.05	0.10	0.25	0.50	0.75	1.00
0.50	187 / 60 / 13	194 / 69 / 10	201 / 79 / 8	264 / 86 / 7	252 / 59 / 3	164 / 10 / 0	113 / 0 / 0
0.75	205 / 37 / 4	201 / 42 / 5	215 / 45 / 6	219 / 66 / 7	275 / 67 / 7	289 / 55 / 3	265 / 23 / 0
0.90	211 / 30 / 1	206 / 30 / 5	219 / 35 / 1	216 / 46 / 4	215 / 59 / 6	240 / 58 / 4	269 / 49 / 3
1.00	212 / 17 / 3	210 / 18 / 2	205 / 22 / 1	221 / 30 / 2	208 / 47 / 3	212 / 41 / 3	218 / 49 / 4

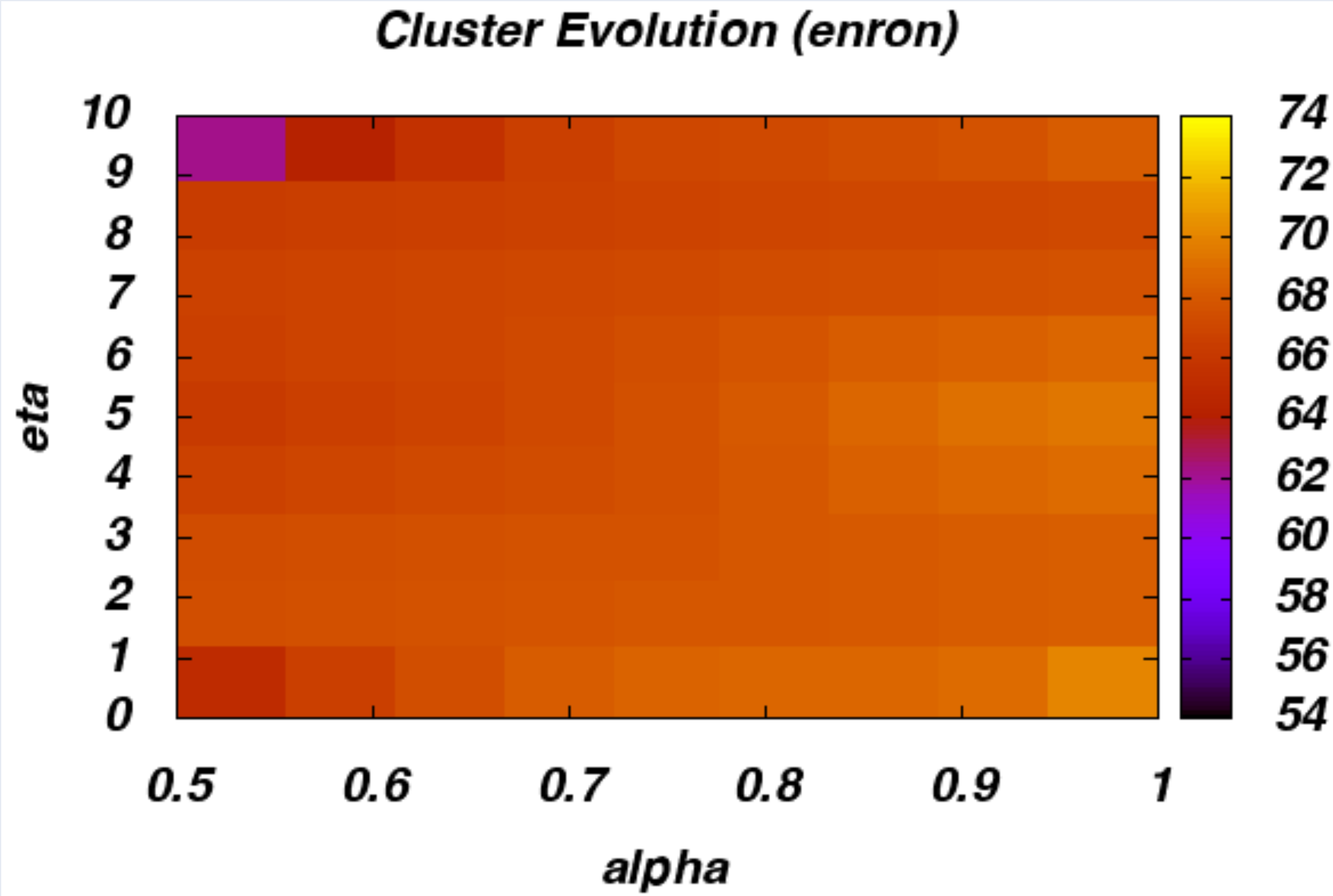
World trade flows nodes

	L / S	L / S	L / S	L / S	L / S	L / S	L / S
α \ η	0.01	0.05	0.10	0.25	0.50	0.75	1.00
0.50	1293 / 4795	1269 / 4768	1361 / 4652	1172 / 4709	585 / 2976	273 / 1432	156 / 718
0.75	906 / 5260	980 / 5173	906 / 5223	1084 / 4935	892 / 4797	610 / 3970	386 / 2782
0.90	803 / 5380	806 / 5304	861 / 5292	843 / 5263	963 / 4986	805 / 4887	640 / 4765
1.00	576 / 5604	599 / 5578	620 / 5555	644 / 5476	702 / 5266	619 / 5202	683 / 4938

Effect on downstream analytics



Effect on downstream analytics



Conclusion

1. There are many ways to pre-process dynamic data
2. Introduced principled parameterized framework
3. Explored how parameters affected various analytics

Take-aways:

1. Varying parameters can uncover structure
2. Different parameters needed to answer different questions
3. Exploring parameters crucial to understand data
4. Need to make explicit what parameters were used in study and why

Thank you

- Sofus A. Macskassy
- Data Scientist, Facebook