Network Analysis Literacy – a socio-informatic approach

Prof. Dr. Katharina A. Zweig

Algorithm Accountability Lab

TU Kaiserslautern

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

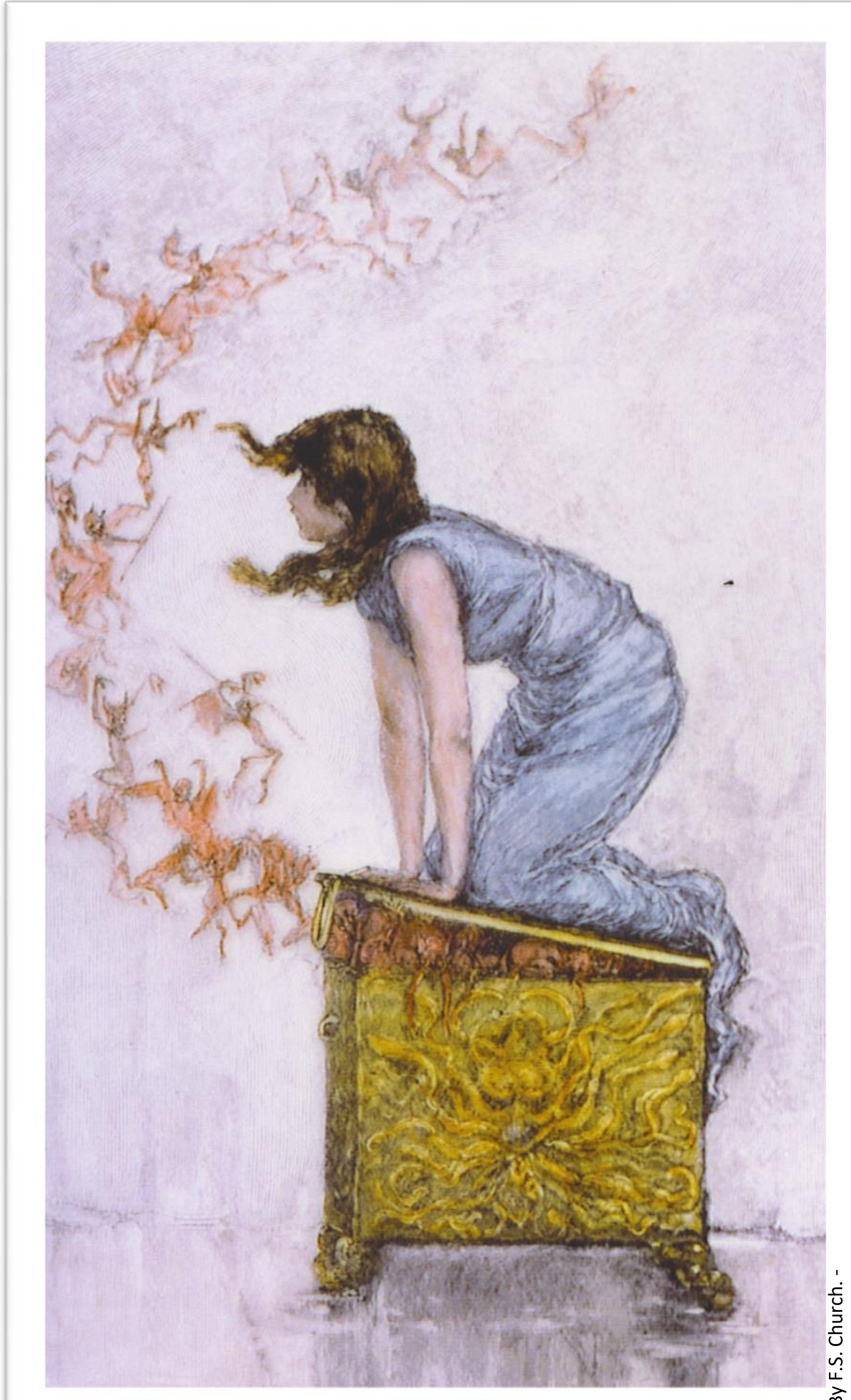
Network Analysis – A basic Toolbox

- Network analysis has become a tool in many sciences:
  - Biology
  - Chemistry
  - Epidemiology
- …but also in many societal contexts:
  - Political advice on, e.g., epidemics prevention
  - Terrorist identification for secret services
- …and maybe soon in many others?
  - China citizen score,
  - credit score based on Facebook,
  - employment based on social media account behavior[1], …

1 https://www.aclu.org/blog/national-security/want-job-password-please?redirect=blog/technology-and-liberty/want-job-password-please

I think we have opened Pandora‘s Box

A drama in three acts



By F.S. Church. - http://prb.livejournal.com/35233.htmlhttp://www.mitchellteachers.org/WorldHistory/AncientGreece/DiscoveringReferencestoGreekMythology.htm, Public Domain, https://commons.wikimedia.org/w/index.php?curid=17344549

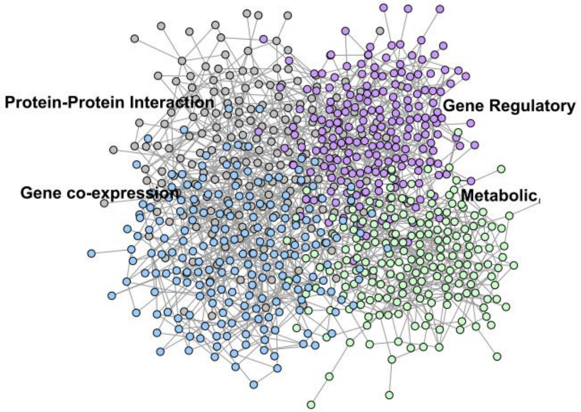
A new look at Centrality Indices

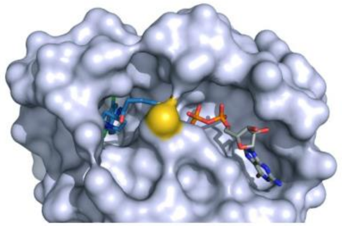
Transferred to multiplex networks

(work with Sude Tavassoli)

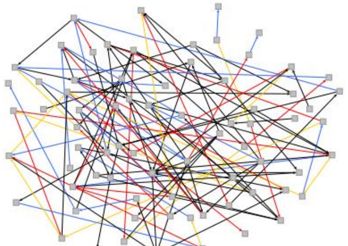
The usefulness of Centrality Measures in Multiplex Networks

- Analyzing flow processes in multiplex networks such as epidemic transmission in Transportation networks [2, 4].
- Identifying cancer drivers in Biological networks using the representation of protein-protein interaction, gene regulation, co-expression, and metabolic network in a multiplex network [1].
- Analyzing leading drivers in Terrorist networks, where for instance, the importance of a node in “communication” layer is affected by the importance of the node in “trust” layer [6].





img: UCSF News Center



Technische Universität Kaiserslautern

So, we could use …

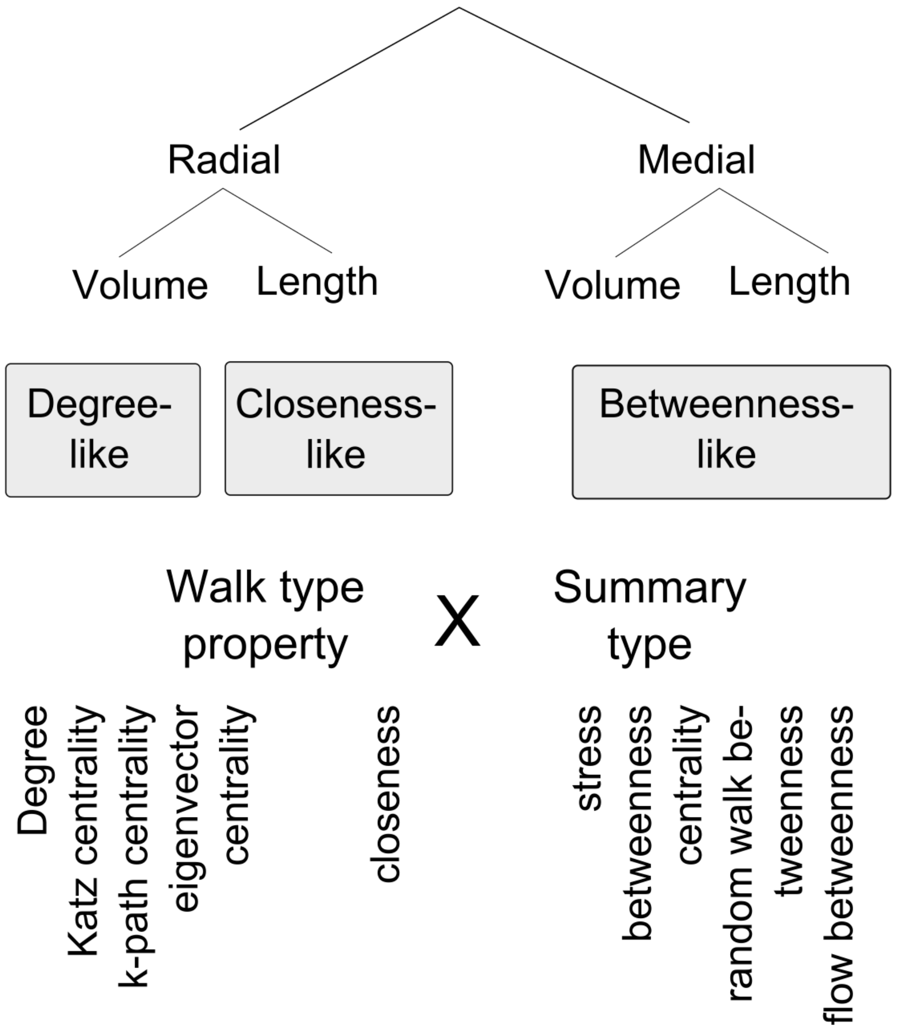
$$C_B(v) = \sum_{s,t \neq v} \frac{\delta_{s,t}(v)}{\delta_{s,t}}$$

1. Act: Wait-wait-wait: Centralities?

Categorizations of Centrality Indices

Borgatti and Everett, 2006

- 1. dimension: walk type?
- 2. dimension: Volume measures (number of paths satisfying some constraint – degree) vs. length measures (counting paths regarding their lengths –closeness)
- 3. dimension: Radial measures (for nodes on the end of paths) vs. medial measures: counting how often a node is on a set of paths.
- 4. dimension: summary type (sum, average, median, …)



Categorizations of Centrality Indices

Borgatti, 2005

- Centrality index is tied to a model of the network flow with certain characteristics:
  - Path type;
  - Serial or parallel diffusion;
  - Divisible, copyable or indivisible good.
- For the matching network flow, it gives the **likelihood of a node of being used**

Weisberg‘s Definition of a Model: Structure + Construal

- Weisberg (2013) argues that models are composed of two things:
  - Their structure
  - A *construal*, the modeler‘s interpretation of the structure.
    - *Assignments* define the *analogy* between the model‘s components and the real-world, target system. E.g.: in social network analysis, nodes represent human actors and edges represent their relationships.
    - *Intended scope*: most modelers have a specific application of the model in mind (but it is not often made explicit)
    - *Fidelity criteria:* standards by which the modeler evaluates the „goodness of fit“ of his or her model to the real-world target system. This can be very different from case to case.

Hidden Assumptions in Betweenness Centrality

Hmmm…

Inherently serial, probably indivisible

Only shortest paths

Okay, that's an approximation, right?

$$C_B(v) = \sum_{s,t \neq v} \frac{\delta_{s,t}(v)}{\delta_{s,t}}$$

And you know that every pair s,t contributes d(s,t)-1 to the total betweenness centrality?

Uoh…

All pairs of nodes want to communicate with the same frequency/intensity

Dorn et al., 2012
Zweig, 2016

Can we use betweenness centrality? Two models need to apply

- Structure I: a model of a network flow
  - Shortest paths, pair-wise interaction with same freq., …
- Construal I:
  - Assignment: real-world flow resembles model
  - Intended scope: flows that are approximated by the model
  - Fidelity criteria ??
- Structure II: most important node is the one used most often expectedly
- Construal II:
  - Assignment: real-world importance to centrality index value
  - Intended scope: when applicable to idea of importance
  - Fidelity criterion: ground truth


Just
xD
SAY
NO

Hidden Assumptions in Betweenness Centrality

Hmmm...

Inherently serial, probably indivisible

Only shortest paths

Okay, that's an approximation, right?

$$C_B(v) = \sum_{s,t \neq v} \frac{\delta_{s,t}(v)}{\delta_{s,t}}$$

And you know that every pair s,t contributes d(s,t)-1 to the total betweenness centrality?

Uoh...

All pairs of nodes want to communicate with the same frequency/intensity

You win. Let's do degree centrality.
We certainly know what that means!

Dorn et al., 2012
Zweig, 2016

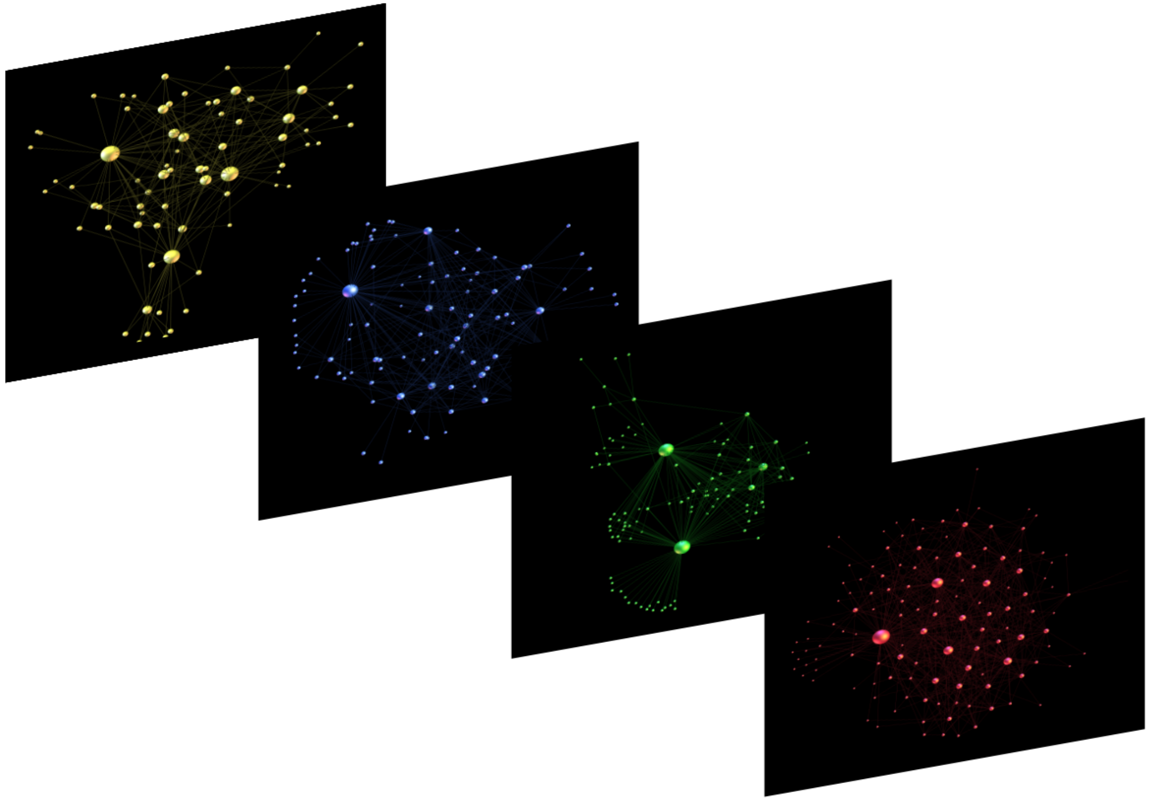
2nd act: Some results

Degree Centrality in Multiplex Networks

Degree Centrality as the simplest index in Multiplex networks

Don‘t forget to normalize!

- A network with $|\mathcal{L}|$ layers $\mathcal{L} = \{L_1, L_2, \cdots, L_{|L|}\}$ where each layer l_i is a simple graph comprised of a set of V_i nodes and $E_i \subseteq V_i \times V_i$ edges.
- A set of nodes are common: $V^* = \bigcap_{i=1}^{|L|} V_i$.
- The degree $deg_i(v)$ of any node v is defined as the number of edges connected to the node v in layer L_i.
- The result of ranking is from position 1 to position $|V^*|$.



Different modeling decisions

The normalization strategies

NormMethod 1, for layer L_i takes $deg_i(v)$ for all $v \in V^*$ and normalizes it with the minimum and maximum values in the set of common nodes. This results in a vector of normalized indices of $[0,1]$ for layer L_i.

$$\mathcal{C}_1(v,i) = \frac{deg_i(v) - min\{deg_i(v)|v \in V^*\}}{max\{deg_i(v)|v \in V^*\} - min\{deg_i(v)|v \in V^*\}}$$

NormMethod 2 is similar to the last method but the normalization is done using the minimum and maximum values in the set of all nodes (V_i) in layer L_i.

$$\mathcal{C}_2(v,i) = \frac{deg_i(v) - min\{deg_i(v)|v \in V_i\}}{max\{deg_i(v)|v \in V_i\} - min\{deg_i(v)|v \in V_i\}}$$

Technische Universität Kaiserslautern

Beautiful, what about aggregation? Most would either use the sum, average, minimum, or maximum degree of one node over all layers.

Tavassoli & Zweig, 2016

Well, I know an operator which can do all of that!

The normalization strategies...

NormMethod 3 uses the results by *NormMethod 2* and multiplies them with the fraction of the maximum degree in layer L_i and the maximum degree among all nodes in all $|\mathcal{L}|$ layers. This results in a vector of indices of nodes ($v \in V_i$) between $[0, \frac{max\{deg_i(v)|v \in V_i\}}{max\{deg_i(v)|v \in \bigcup V_j, 1 \le i \le |\mathcal{L}|\}}]$.

$$\mathcal{C}_3(v,i) = \mathcal{C}_2(v) \cdot \left(\frac{max\{deg_i(v)|v \in V_i\}}{max\{deg_i(v)|v \in \bigcup V_j, i \in [1,\ldots,|\mathcal{L}|]\}} \right)$$

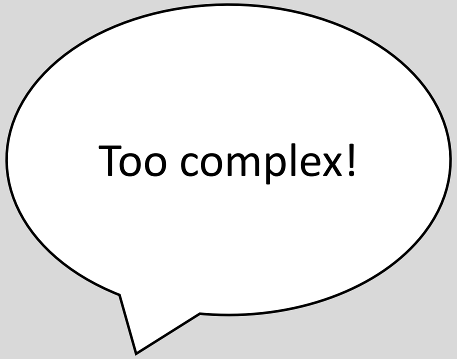
NormMethod 4 for each layer, we rank the nodes non-increasingly by their degree $deg_i(v)$ and obtain $r_i(v)$. This is then normalized by n_i.

$$\mathcal{C}_4(v,i) = \frac{r_i(v)}{n_i}$$

Technische Universität Kaiserslautern

Different modeling decisions

The aggregation strategies



Maximum Entropy Ordered Weighted Averaging (MEOWA) operator (denoted by λ) creates a single number based on the vector of a node's $|\mathcal{L}|$ normalized degrees as follows:

$$\lambda(C_x(v,1), C_x(v,2), \cdots, C_x(v,|\mathcal{L}|)) = \sum_j w_j \; d_j(v)$$

where $D = (b_1, b_2, ..., b_{|\mathcal{L}|})$ is the non-increasingly sorted vector of the normalized degrees, and w is a weight vector. The weight vector is obtained using the following function based on a parameter β [5]:

$$w_i = \frac{e^{\beta \frac{n-i}{n-1}}}{\sum_{j=1}^{n} e^{\beta \frac{n-j}{n-1}}}.$$

Wait-wait-wait: It‘s Fuzzy!

Manchester‘s degree in 4 airlines

1
12
5
5

Sort it!

1
5
5
12

Choose a β-value

Very low β

1
0
0
0

= 1

β=0

0,25
0,25
0,25
0,25

= 5,75

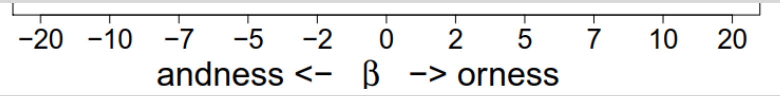
Very high β

0
0
0
1

= 12

For historical reasons, we speak of „high andness“ and „high orness“:

We either require the degrees of a node to be high on ALL layers („and“) or on at least one („or“).



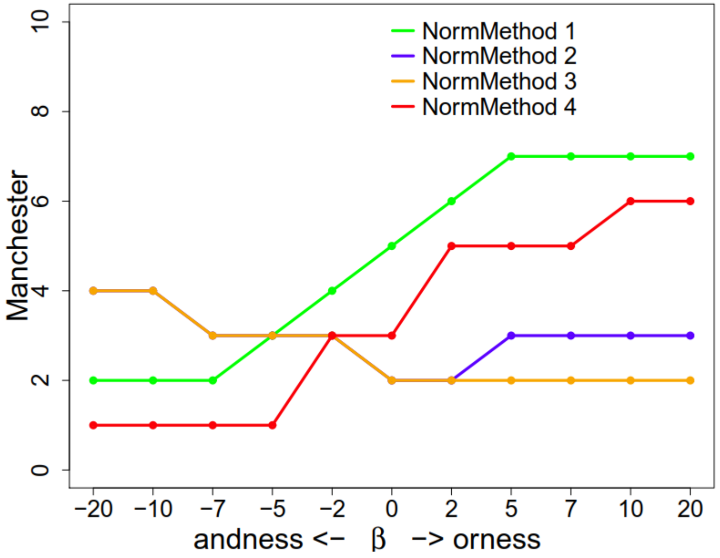
Zadeh, 1965

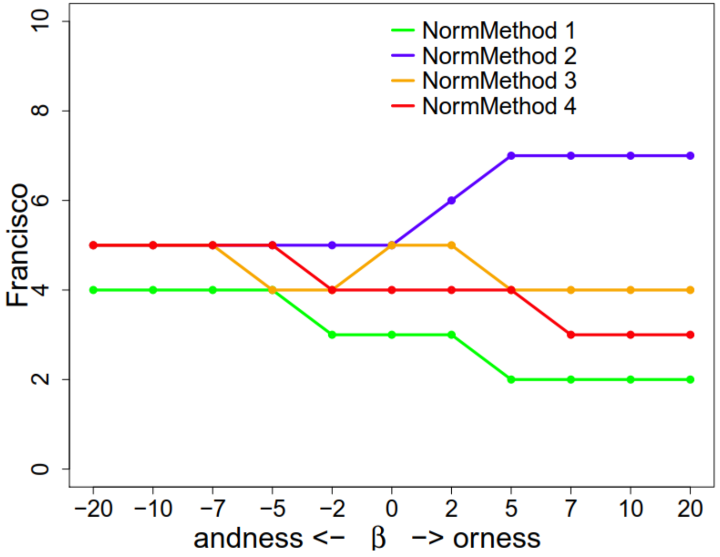
Okay, what are the results?

European airlines data set

A network comprised of four layers of airlines: Air Berlin, Easyjet, Lufthansa, and Ryan air. The order varies from 75 to 128 among four layers [2]. 9 nodes are common among the four layers.

Okay. Both, Manchester and Francisco can be third most central – or third least central. Can we quantify this?





Properties	Air-Berlin	Easyjet	Lufthansa	Ryanair
$\lvert V_i \rvert$	75	99	106	128
$\lvert E_i \rvert$	239	347	244	601
$\max_{v\in V_i}\{deg(v)\}$	37	67	78	85
$\max_{v\in V^*}\{deg(v)\}$	26	17	5	28
$\min_{v\in V_i}\{deg(v)\}$	1	1	1	1
$\min_{v\in V^*}\{deg(v)\}$	1	2	1	5

E.G.,

$deg(Manchester)$: 1, 12, 5, 5 $\rightarrow \mathcal{C}_1(v)$: 0, 0.667, $\boxed{1}$, 0

$\mathcal{C}_2(v)$: 0, $\frac{11}{66}$, $\frac{4}{77}$, $\frac{4}{84}$ $\rightarrow$ 0, $\boxed{0.167}$, 0.052, 0.048

$\mathcal{C}_3(v)$: $\mathcal{C}_2(v) \cdot (\frac{37}{85}, \frac{67}{85}, \frac{78}{85}, \frac{85}{85}) \rightarrow$ 0, $\boxed{0.131}$, 0.048, 0.048

$\mathcal{C}_4(v)$: 0.093, 0.818, $\boxed{0.887}$, 0.461

$deg(Francisco)$: 12, 5, 1, 15 $\rightarrow \mathcal{C}_1(v)$: $\boxed{0.44}$, 0.2, 0, 0.435

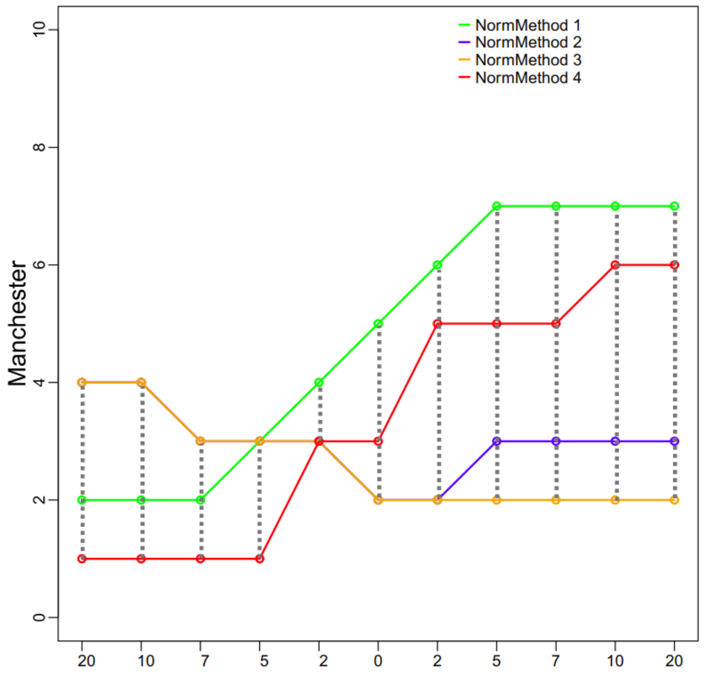
$\mathcal{C}_2(v)$: $\boxed{0.306}$, 0.061, 0, 0.167

$\mathcal{C}_3(v)$: 0.133, 0.048, 0, $\boxed{0.167}$

$\mathcal{C}_4(v)$: $\boxed{0.833}$, 0.611, 0.184, 0.789

Example:

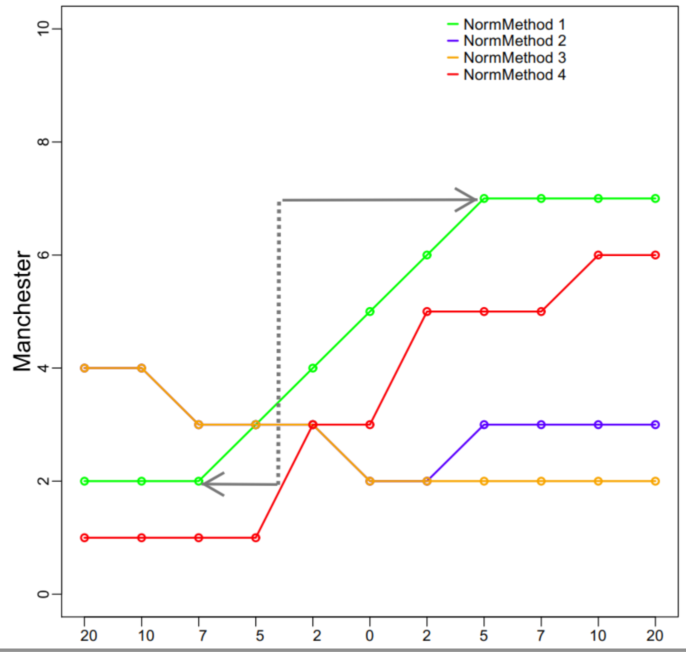
$\Delta norm(Manchester) := max\{3, 3, 2, 2, 1, 3, 4, 5, 5, 5, 5\} = 5$



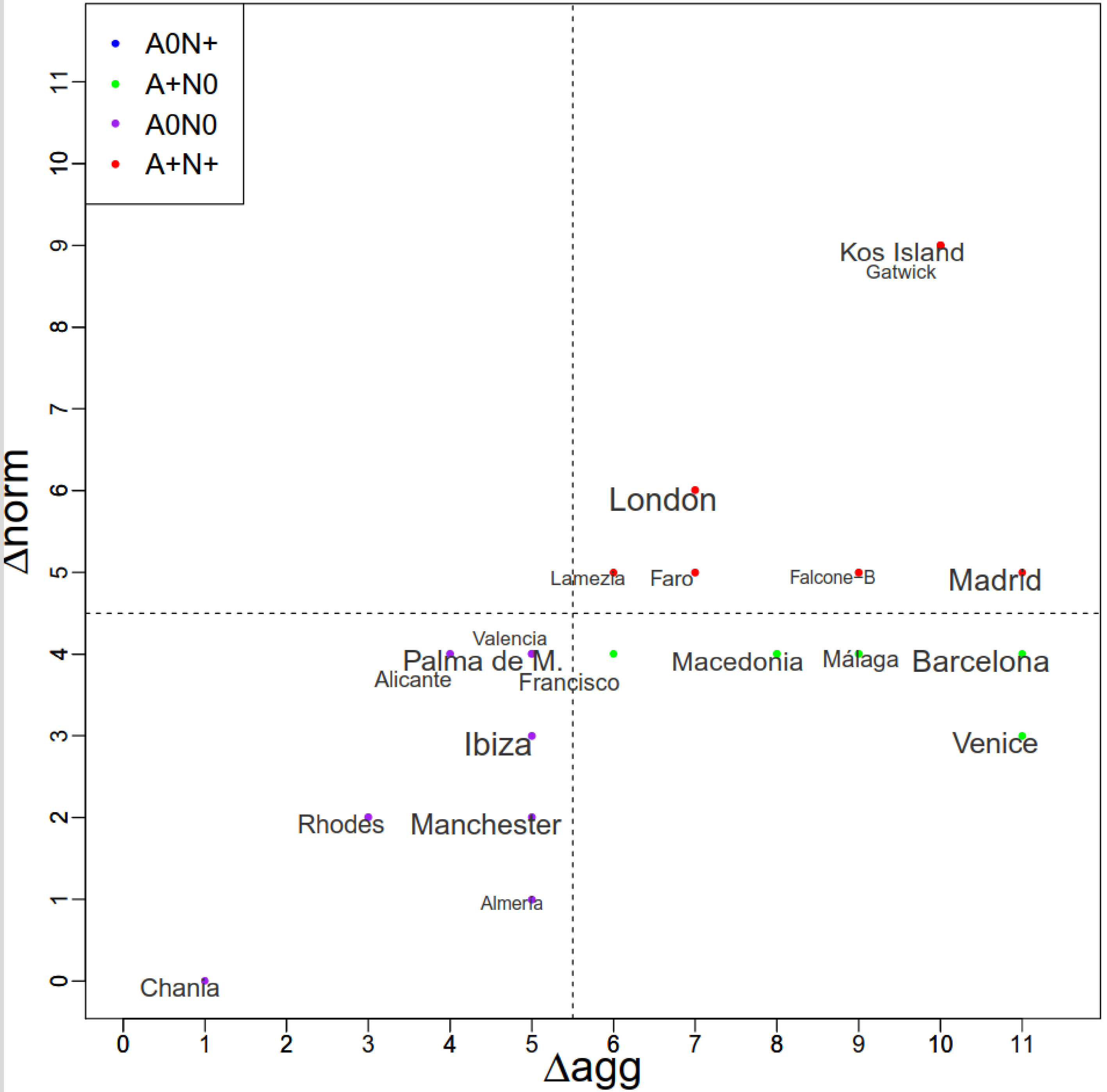
If we leave out Lufthansa, there are 20 common nodes between the other three airlines.

Example:

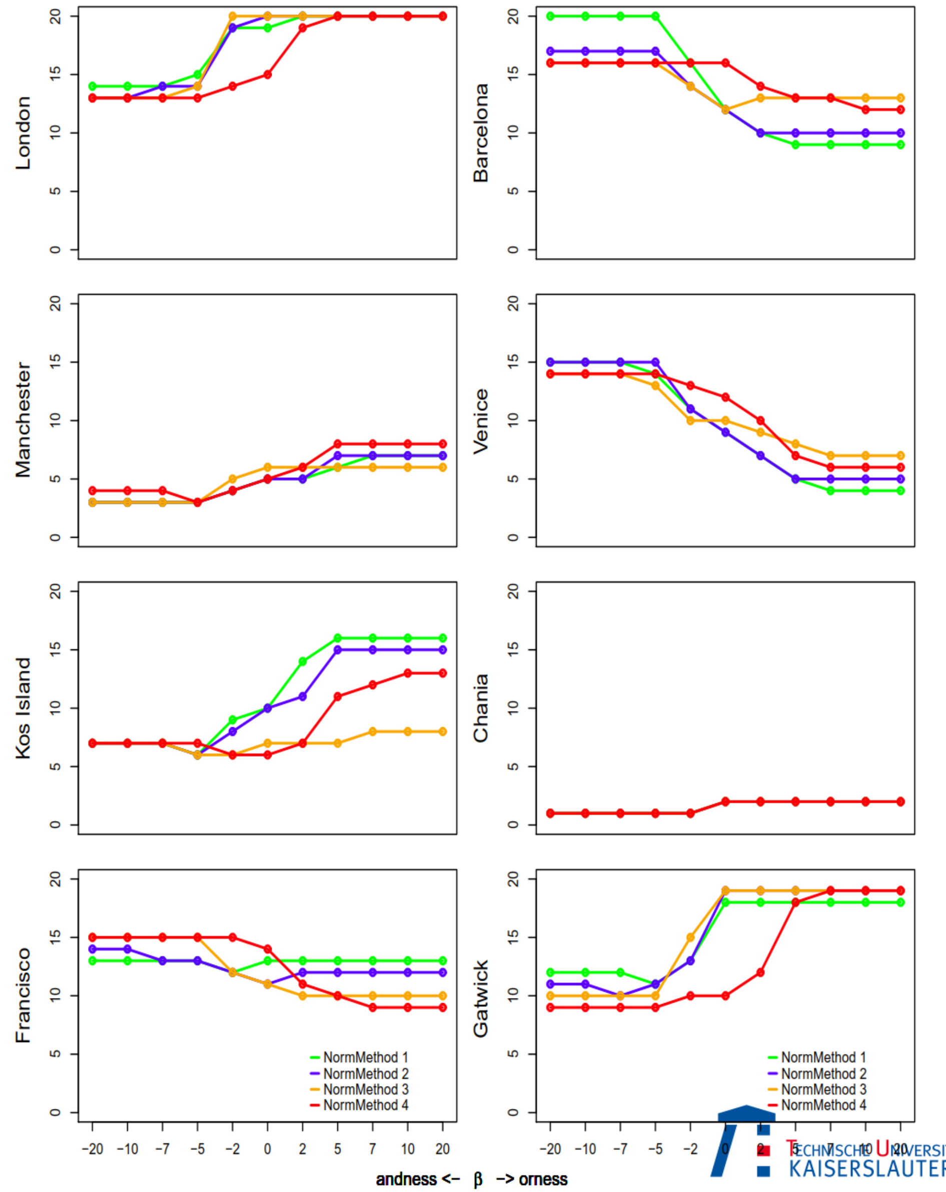
$\Delta agg(Manchester) := max\{5, 2, 2, 5\} = 5$



Let's plot this for all nodes - wait, there are only 9 of them.



he aggregation scenario, then we have 20 common



TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

A network comprised of three layers of seeking advice, ... and having a friendship outside the firm among 71 attorneys [8].

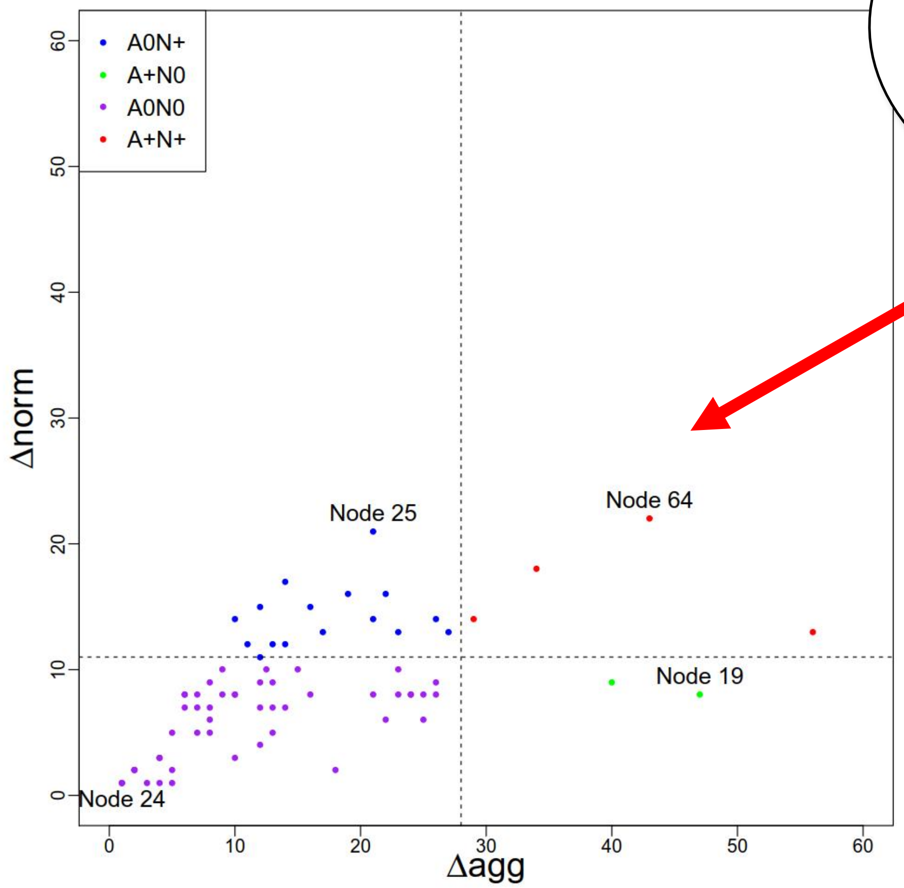


Figure: The sensitivity of 71 nodes to the choices of different aggregation strategies (Δagg) and the different normalization methods ($\Delta norm$).

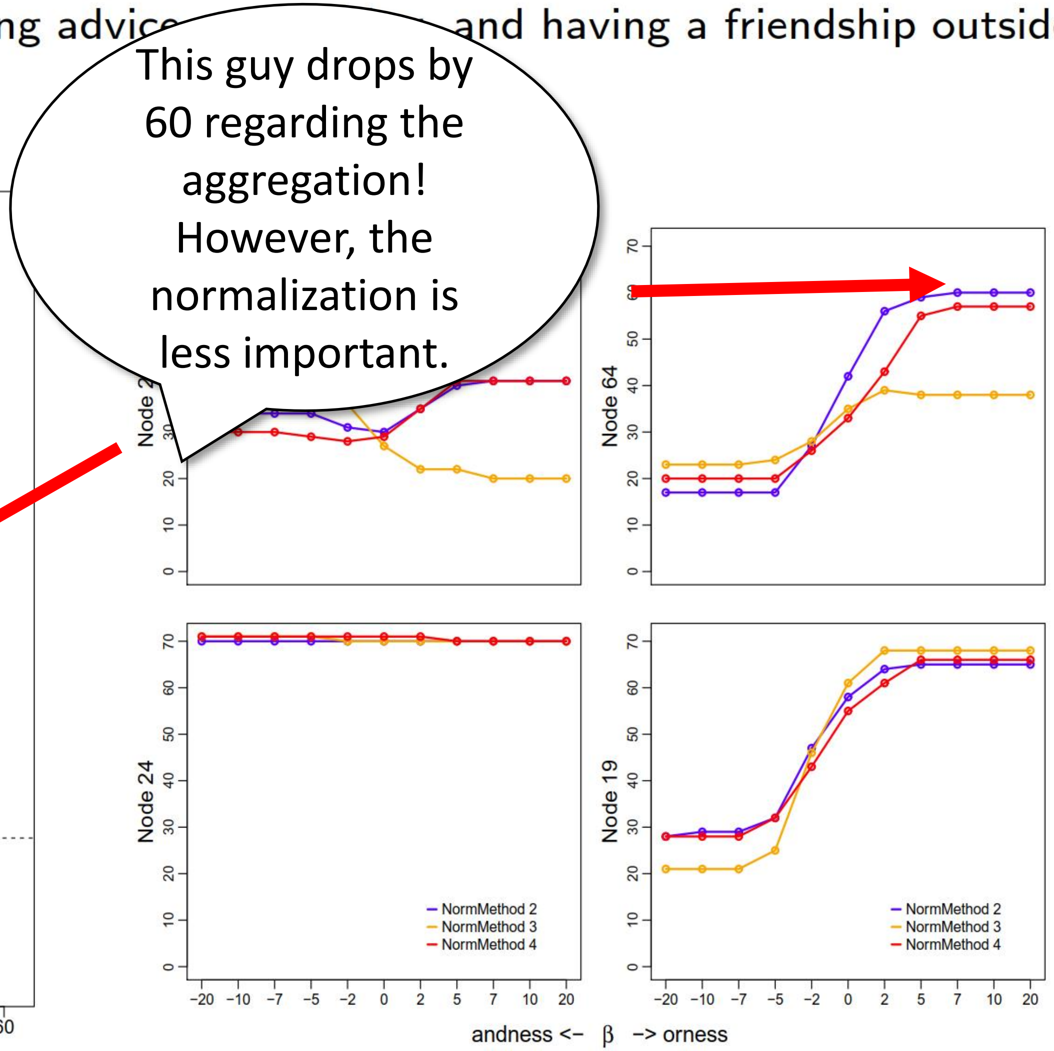


Figure: The rankings obtained using the different aggregation strategies (using the β parameter) for the aggregation of the results of three layers.

Tweets dataset

A network comprised of four layers representing dif

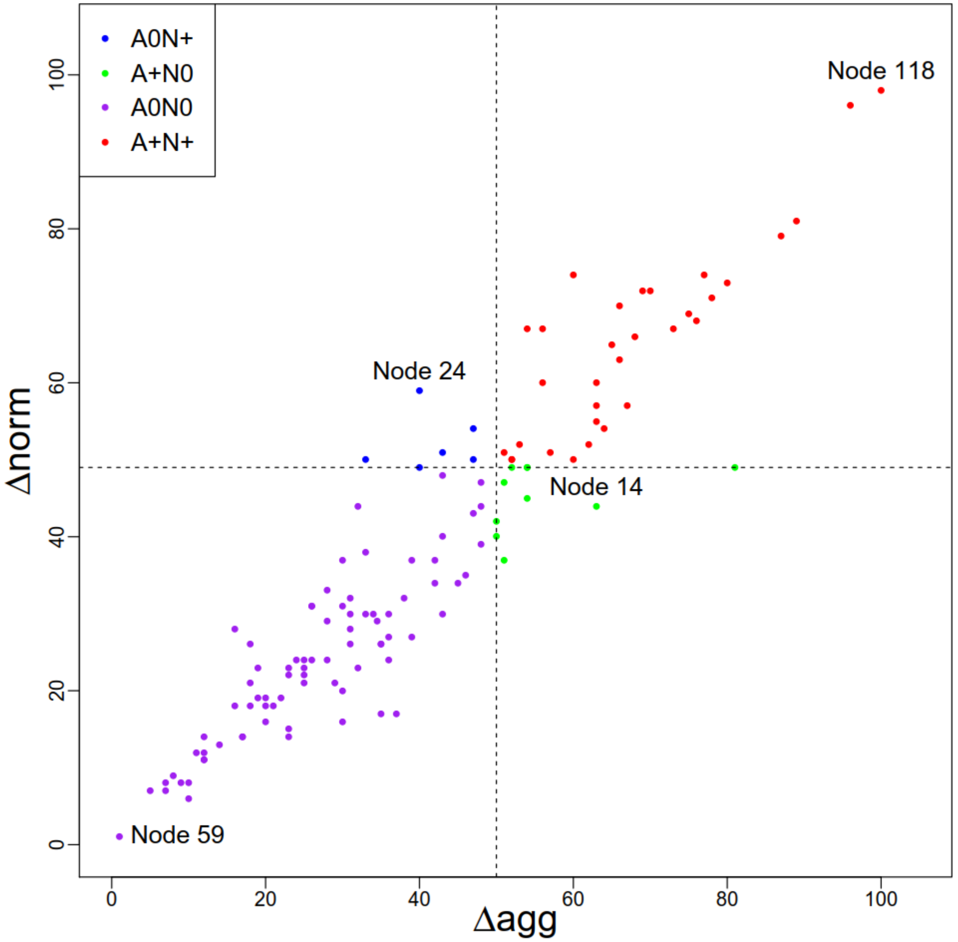
oncerning the

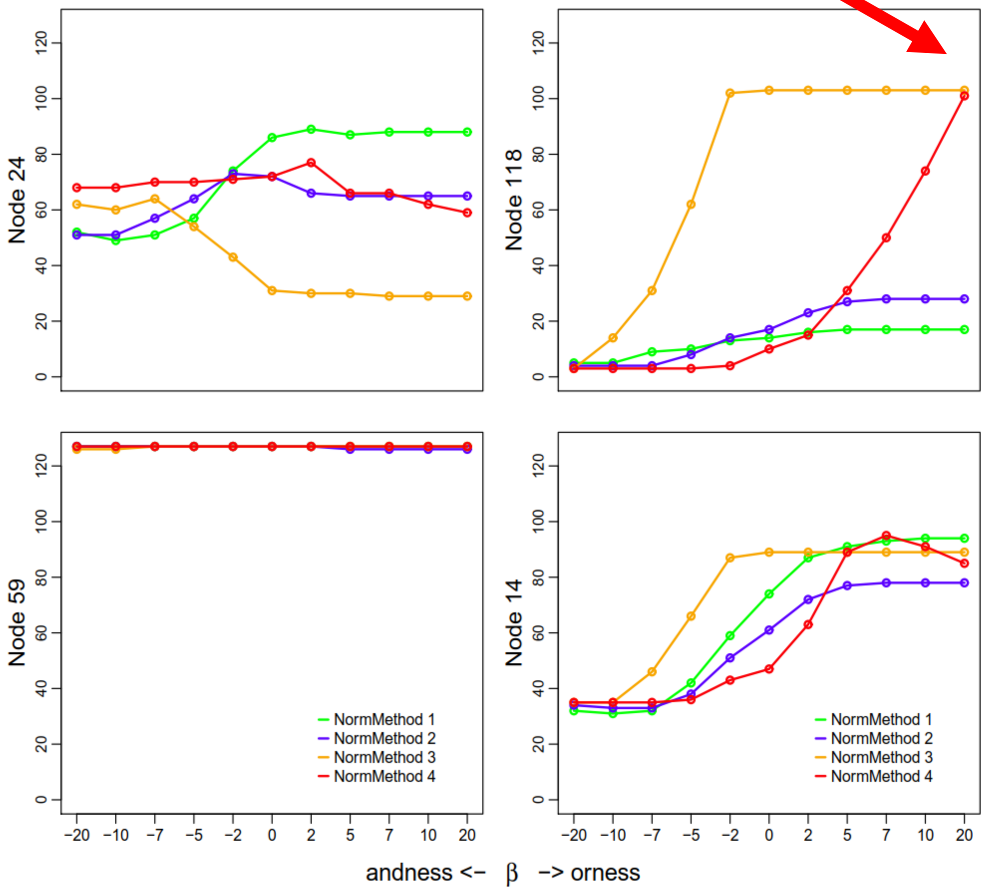
"Higgs Boson": *mentioning*, *replying* to the tweets, re

ner users, plus

the social network of followers/followees [3].

This guy drops by 100? Out of 127? Puh. And it is sensitive to both, normal and agg!





Update

- Betweenness centrality and other centrality indices make assumptions that are not likely to be true in real-world scenarios
- But even the degree centrality is hard to interpret.
  - Normalization necessary
  - Aggregation necessary
  - Different sensitivities

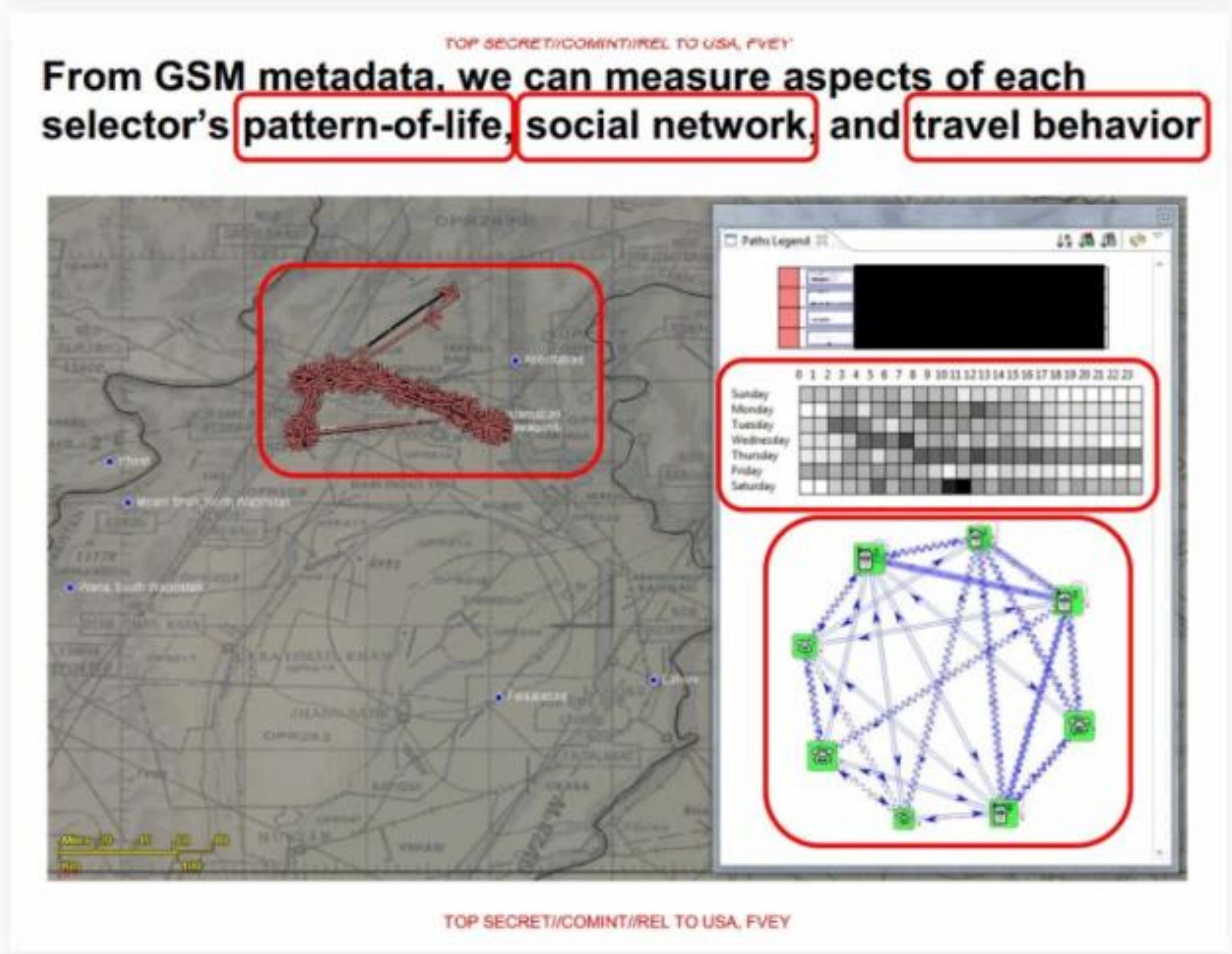
3rd act:
Literacy and Accountability

Network analysis literacy

- Network analysis was used to convey to politicians whom to take care of in HIV and other sexual disease spreadings (Butts, 2009)
- It‘s been used to discredit a climate modeling scientist (Zweig, 2016)
- Network analysis is used to find terrorists…

„Rural politics“ („Die Dorfpolitiker“), Friedrich Friedländer

Capturing terrorists with network analysis



Terrorist identification SKYNET

TOP SECRET//COMINT//REL TO USA, FVEY

We've been experimenting with several error metrics on both small and large test sets

Training Data	Classifier	Features	100k Test Selectors: False Alarm Rate at 50% Miss Rate	100k Test Selectors: Mean Reciprocal Rank	55M Test Selectors: Tasked Selectors in Top 500	55M Test Selectors: Tasked Selectors in Top 100
None	Random	None	50%	1/23k (simulated)	0.64 (active/Pak)	0.13 (active/Pak)
Known Couriers	Centroid	All	20%	1/18k		
		Outgoing	43%	1/27k		
	Random Forest	Outgoing	0.18%	1/9.9	5	1
+ Anchory Selectors	Random Forest	Outgoing	0.008%	1/14	21	6

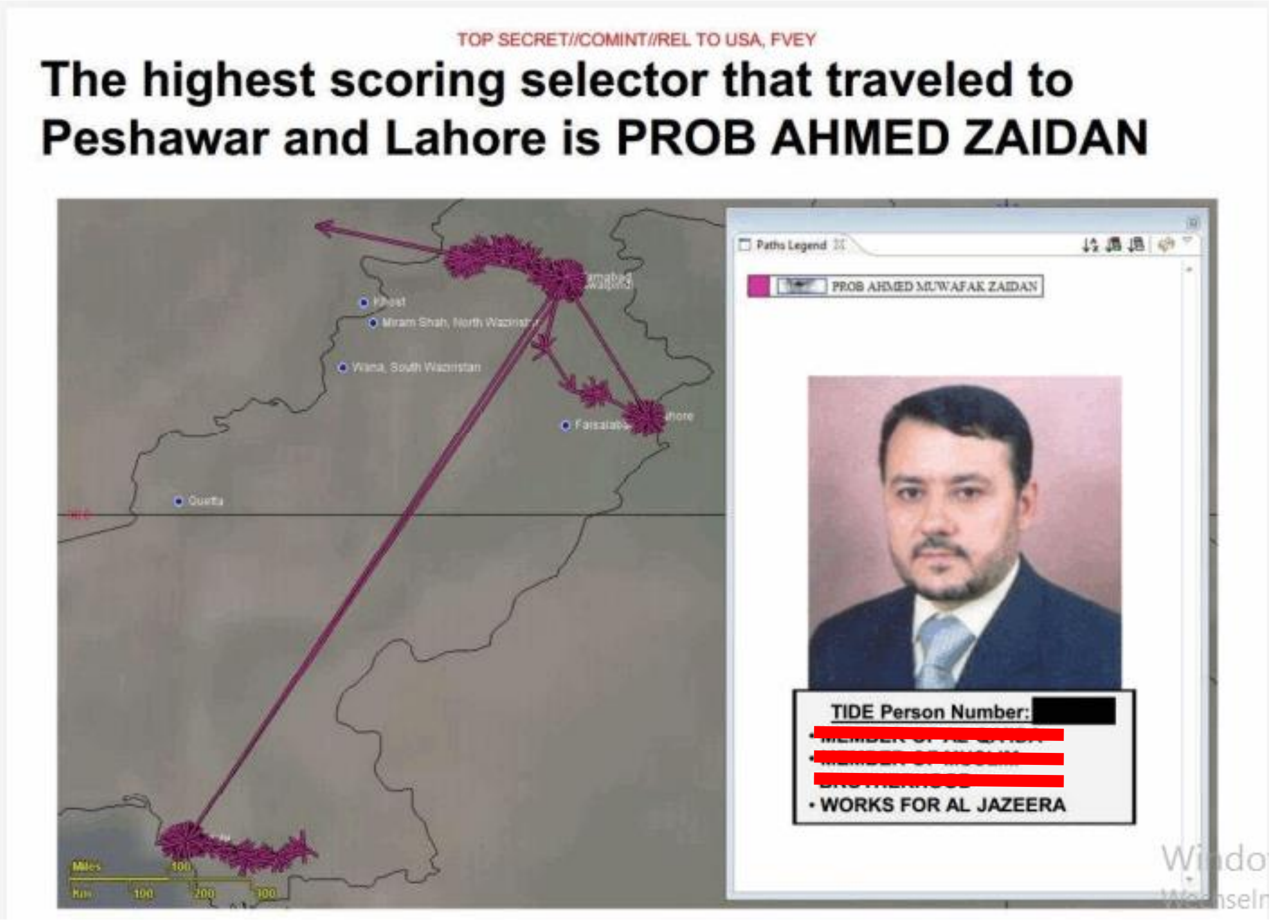
Random Forest trained on Known Couriers + Anchory Selectors:

- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

TOP SECRET//COMINT//REL TO USA, FVEY

https://theintercept.com/document/2015/05/08/skynet-courier/
https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/

Top-“terrorist courier“ is…

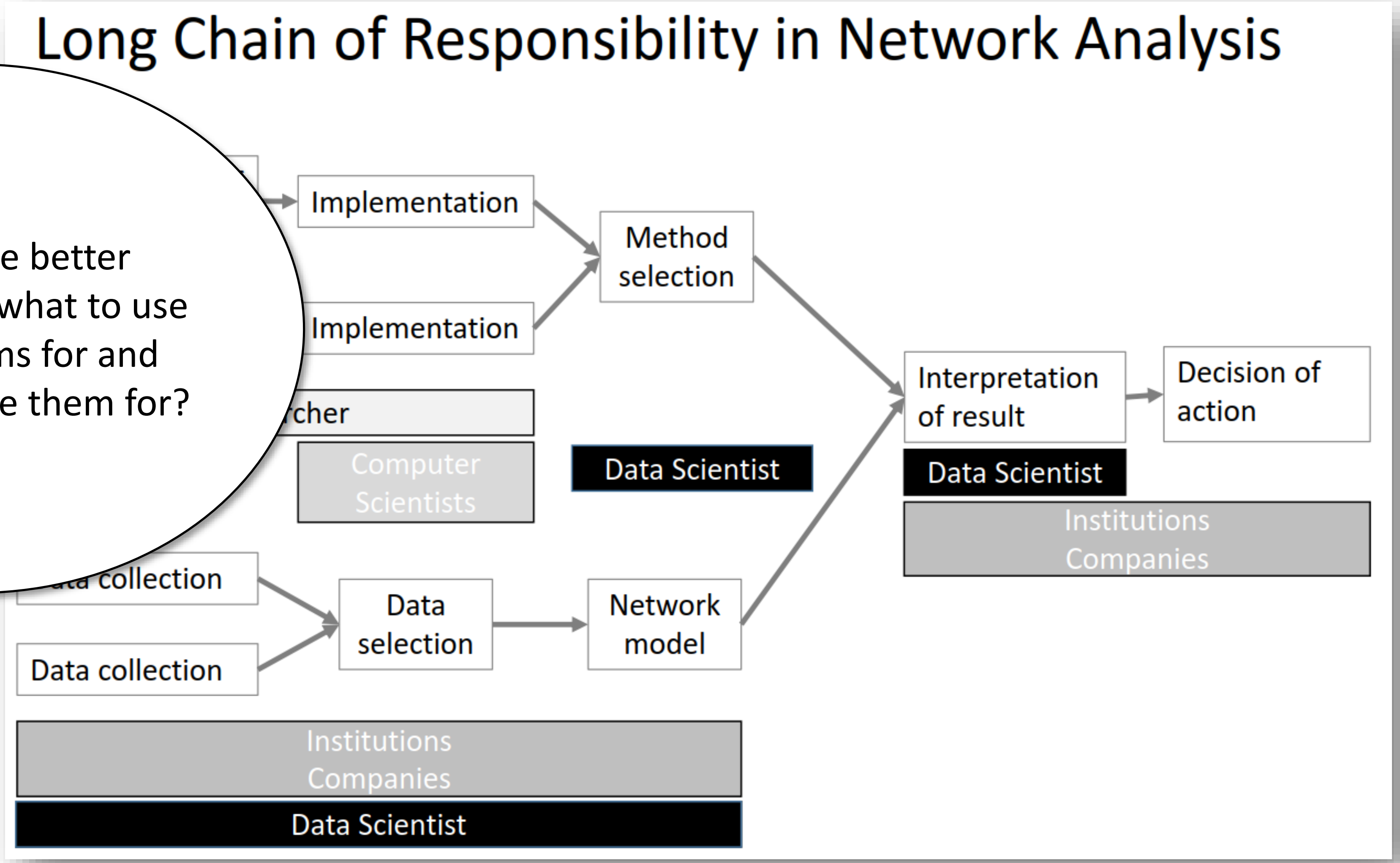


Network Analysis Literacy

- Networks are models of real-life systems.
- A measure is essentially a *model* of what you think the edges mean and how they are used.
- To make interpretations of the results, both models (network/measure) need to match your research question.

Algorithm Accountability

ALGORITHM WATCH
algorithmwatch.org



How can we better communicate what to use our algorithms for and what not to use them for?

Gründung von „Algorithm Watch“

ALGORITHM WATCH

Lorena Jaume-Palasí, Mitarbeiterin im iRights.Lab

Lorenz Matzat, Datenjournalist der 1. Stunde, Gründer von lokaler.de, Grimme-Preis-Träger

Matthias Spielkamp, Gründer von iRights.info, ebenfalls Grimme-Preis-Träger, Vorstandsmitglied von Reporter ohne Grenzen.

Prof. Dr. K.A. Zweig, Junior Fellow der Gesellschaft für Informatik, Digitaler Kopf 2014, TU Kaiserslautern

Literature

- [Borg2005] Stephen P. Borgatti. Centrality and network flow. Social Networks, 27:55–71, 2005.
- [Borg2006] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. Social Networks, 28:466–484, 2006.
- [Butts2009] Carter T. Butts. Revisiting the foundations of network analysis. Science, 325(5939):414–416, 2009
- [Zadeh1965] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3):338–353
- [Zweig2016] Zweig, K.A.: Network analysis literacy, Springer Verlag Wien, 2016