# Some topics in network science

Mark C. Wilson
University of Auckland

Department of Mathematics and Statistics
UMass Amherst
2017-11-27

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.
- Networks are ubiquitous. A possible pitfall is not to consider the meaning of the relation being graphed.

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.
- Networks are ubiquitous. A possible pitfall is not to consider the meaning of the relation being graphed.
- Will discuss 3 examples today, exemplifying:

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.
- Networks are ubiquitous. A possible pitfall is not to consider the meaning of the relation being graphed.
- Will discuss 3 examples today, exemplifying:
  - Construction of a network from real data, some basic network science tools.

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.
- Networks are ubiquitous. A possible pitfall is not to consider the meaning of the relation being graphed.
- Will discuss 3 examples today, exemplifying:
  - Construction of a network from real data, some basic network science tools.
  - Social learning and "wisdom of crowds".

# Basics

- A network is a finite graph (sometimes directed) $G = (V, E)$.
- Network science deals with real-world networks, often varying with time.
- The main questions relate to network *structure* and *evolution* over time.
- Networks are ubiquitous. A possible pitfall is not to consider the meaning of the relation being graphed.
- Will discuss 3 examples today, exemplifying:
  - Construction of a network from real data, some basic network science tools.
  - Social learning and "wisdom of crowds".
  - Structural balance in signed networks.

# Citation networks

- Here nodes are documents and (directed) edges are formed when one cites another.

# Citation networks

- Here nodes are documents and (directed) edges are formed when one cites another.
- We are all familiar with the citation network of scientific papers.

# Citation networks

- Here nodes are documents and (directed) edges are formed when one cites another.
- We are all familiar with the citation network of scientific papers.
- The citation network of Supreme Court (USA) opinions has been analysed.

# Citation networks

- Here nodes are documents and (directed) edges are formed when one cites another.
- We are all familiar with the citation network of scientific papers.
- The citation network of Supreme Court (USA) opinions has been analysed.
- A relatively new example is the corpus of legislative documents (acts, regulations, case law). My PhD student Neda Sakhaee and I looked at New Zealand Acts of Parliament (in progress).

## Citation networks

- Here nodes are documents and (directed) edges are formed when one cites another.
- We are all familiar with the citation network of scientific papers.
- The citation network of Supreme Court (USA) opinions has been analysed.
- A relatively new example is the corpus of legislative documents (acts, regulations, case law). My PhD student Neda Sakhaee and I looked at New Zealand Acts of Parliament (in progress).
- Basic questions: what is the network structure? how does it evolve? which are the "most important/influential" documents? do they cluster?

# Notes on real data

- Network analyses can be sensitive to missing data, because we consider not only direct connections, but those at greater distance. Errors can propagate.

# Notes on real data

- Network analyses can be sensitive to missing data, because we consider not only direct connections, but those at greater distance. Errors can propagate.
- Getting hold of real data can be very hard. For scientific citations, the for-profit companies will not share it reasonably. Luckily all NZ laws are available online.

# Notes on real data

- Network analyses can be sensitive to missing data, because we consider not only direct connections, but those at greater distance. Errors can propagate.
- Getting hold of real data can be very hard. For scientific citations, the for-profit companies will not share it reasonably. Luckily all NZ laws are available online.
- Even if it is open, it may not be machine-readable. We have spent much time processing data automatically and manually. This involves large-scale OCR of documents, which is noisy.

# Notes on real data

- Network analyses can be sensitive to missing data, because we consider not only direct connections, but those at greater distance. Errors can propagate.
- Getting hold of real data can be very hard. For scientific citations, the for-profit companies will not share it reasonably. Luckily all NZ laws are available online.
- Even if it is open, it may not be machine-readable. We have spent much time processing data automatically and manually. This involves large-scale OCR of documents, which is noisy.
- Luckily there is a master title list of laws, and the NZ government makes *current* laws available in XML format.

NEW ZEALAND.

TRICESIMO NONO

# VICTORIÆ REGINÆ.

No. LXXIV.

ANALYSIS.



AN ACT to provide for the Collection by means of Stamps of Fees payable in the various Departments of the Public Service. Title.

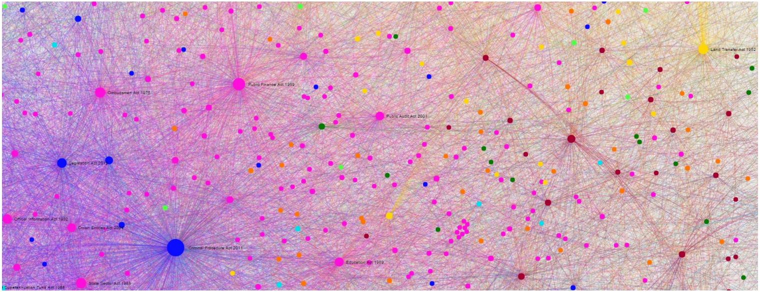
[21st *October*, 1875.]

BE IT ENACTED by the General Assembly of New Zealand in Parliament assembled, and by the authority of the same, as follows:—

**1.** The Short Title of this Act shall be "The Stamp Fee Act, 1875." Short Title.

```
10.
Part I. of "Stamp Act, 1875," to be read al part of this Act. '


AN ACT to provide for the Collection by means of Titlo. Stamps of Fees payable in the
various Depart≠ments of the Public Service. [21St October, 1875.]
B
E IT ENACTED by the General Assembly of New Zealand in
Parliament assembled, and by the a:uthority of the same, as

follows:≠
1.
The Short Title of this Act shall be "The Stamp Fee Act, SIJOl't Title. 1875."

2.
"The Supreme Court and Registration Offices Fees Act, 1866," Repeal. is hereby repealed.

3.
The Governor in Council may, by notice published in the New Governor may fix Zealand
Gazette, direct that after the time specified in such notice Atimte,f∞tor bringit~g


d t' f fi l' .c th t' b . C lD opera Ion
all     or any 0 f the u les ees nes or pena tIes J.or e Ime emg in any Department.
payable in money in any Public Department or office connected with
the public service, or to the officers thereof, shall be collected by
means of stamps; and after the time so specified, the duties fees
fines or penalties therein mentioned shall be received by stamps de≠
noting the sums payable and not in money.
4. The Governor in Council may make alter or repeal Regula-Governor may make Hons not
contrary to this Act for the due administration thereof. Regulations.
Sl~pplelnent to the New Zealand Gazette, No. 59, oftlte 2ht OcfollFr, I8i5.

39∞ VICTORIlE. No. 74.

Stamp Fee Act.
Stamps to be im-5. All or any stamps to be used under this Act shall be impressed
preGssed or addh.esivte or adhesive as the Governor from time to time directs.
as overnor Iree B. 6    . d' h" bl'
Stamp to be affixed to ï When any sum comprIse III any suc notICe IS paya e III
orimpress~duponthe respect of a document the stamps denoting such sum shall be affixed
document ill respect    ï
```

# Part of the network – color:community, node size:centrality




There are 10000s of nodes,100000s of edges.

# Importance of nodes

- The value-neutral term is centrality. There are many measures.

# Importance of nodes

- The value-neutral term is centrality. There are many measures.
- Let $A$ be the adjacency matrix of $G$. Then the centrality vector satisfies

# Importance of nodes

- The value-neutral term is centrality. There are many measures.
- Let $A$ be the adjacency matrix of $G$. Then the centrality vector satisfies
  - $C_I$, indegree centrality (more citations, more important) is given by $C_I = A^T\mathbf{1}$.

## Importance of nodes

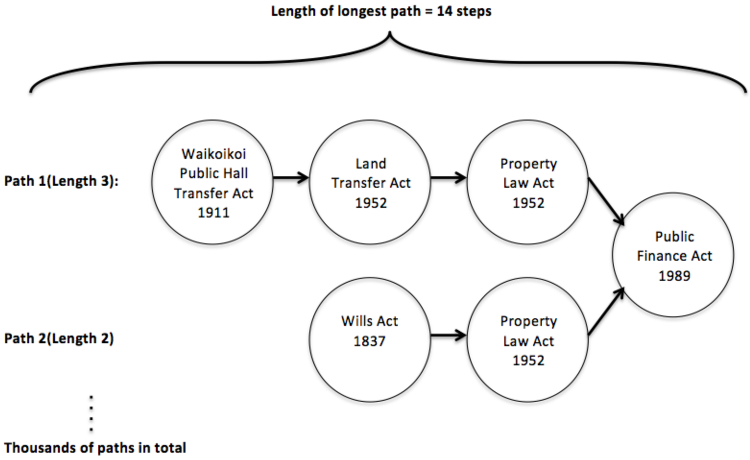
- The value-neutral term is centrality. There are many measures.
- Let $A$ be the adjacency matrix of $G$. Then the centrality vector satisfies
  - $C_I$, indegree centrality (more citations, more important) is given by $C_I = A^T\mathbf{1}$.
  - $C_E$, eigenvalue centrality (similar to Google PageRank, being cited by important nodes makes you important) is given by $A^TC_E = \lambda C_E$ where $\lambda$ is the largest eigenvalue of $A$.

# Importance of nodes

- The value-neutral term is centrality. There are many measures.
- Let $A$ be the adjacency matrix of $G$. Then the centrality vector satisfies
  - $C_I$, indegree centrality (more citations, more important) is given by $C_I = A^T\mathbf{1}$.
  - $C_E$, eigenvalue centrality (similar to Google PageRank, being cited by important nodes makes you important) is given by $A^T C_E = \lambda C_E$ where $\lambda$ is the largest eigenvalue of $A$.
  - $C_K$, Katz centrality (weighted sum of paths to the node, weights decrease exponentially by length) is given by $C_K = ((I - \alpha A^T)^{-1} - I)\mathbf{1}$ for small $\alpha$.

# Importance of nodes

- The value-neutral term is centrality. There are many measures.
- Let $A$ be the adjacency matrix of $G$. Then the centrality vector satisfies
  - $C_I$, indegree centrality (more citations, more important) is given by $C_I = A^T \mathbf{1}$.
  - $C_E$, eigenvalue centrality (similar to Google PageRank, being cited by important nodes makes you important) is given by $A^T C_E = \lambda C_E$ where $\lambda$ is the largest eigenvalue of $A$.
  - $C_K$, Katz centrality (weighted sum of paths to the node, weights decrease exponentially by length) is given by $C_K = ((I - \alpha A^T)^{-1} - I)\mathbf{1}$ for small $\alpha$.
- We found broad agreement between measures on the most important nodes and on the least important, *without any analysis of the content of documents*. This has been corroborated by expert opinion.


Length of longest path = 14 steps
Path 1(Length 3):
Waikoikoi Public Hall Transfer Act 1911
Land Transfer Act 1952
Property Law Act 1952
Public Finance Act 1989
Path 2(Length 2)
Wills Act 1837
Property Law Act 1952
Thousands of paths in total

| Act | Rank | $C_K$ |
|---|---|---|
| Public Finance Act 1989 | 1 | 10.37 |
| Criminal Procedure Act 2011 | 2 | 9.65 |
| Summary Proceedings Act 1957 | 3 | 9.28 |
| State Sector Act 1988 | 4 | 8.85 |
| District Courts Act 1947 | 5 | 7.96 |
| Crimes Act 1961 | 6 | 7.47 |
| Companies Act 1993 | 7 | 7.43 |
| Local Government Act 1974 | 8 | 7.4 |
| Judicature Act 1908 | 9 | 7.1 |
| Privacy Act 1993 | 10 | 6.79 |
| Resource Management Act 1991 | 11 | 6.71 |
| Official Information Act 1982 | 12 | 6.58 |

# Future work

- We aim to study and model evolution of the network over time, but the further back we go, the noisier the data is. The network is becoming denser over time.

# Future work

- We aim to study and model evolution of the network over time, but the further back we go, the noisier the data is. The network is becoming denser over time.
- Interpretation of results is tricky. What is the underlying reason for citation? We aim to correlate changes in network structure with external political and economic events.

# Future work

- We aim to study and model evolution of the network over time, but the further back we go, the noisier the data is. The network is becoming denser over time.
- Interpretation of results is tricky. What is the underlying reason for citation? We aim to correlate changes in network structure with external political and economic events.
- Detect communities (unusually dense subgraphs) - challenging because network is directed.

# Future work

- We aim to study and model evolution of the network over time, but the further back we go, the noisier the data is. The network is becoming denser over time.
- Interpretation of results is tricky. What is the underlying reason for citation? We aim to correlate changes in network structure with external political and economic events.
- Detect communities (unusually dense subgraphs) - challenging because network is directed.
- Comparative studies with other jurisdictions — how much can be read off just from the citation network?

# Future work

- We aim to study and model evolution of the network over time, but the further back we go, the noisier the data is. The network is becoming denser over time.
- Interpretation of results is tricky. What is the underlying reason for citation? We aim to correlate changes in network structure with external political and economic events.
- Detect communities (unusually dense subgraphs) - challenging because network is directed.
- Comparative studies with other jurisdictions — how much can be read off just from the citation network?
- Other layers (regulations, case law) of the network.

# References

- N. Sakhaee, M.C. Wilson, G.Zakeri. *Structural Analysis of Legislation Networks*. Proceedings JURIX 2016.

# References


- N. Sakhaee, M.C. Wilson, G.Zakeri. *Structural Analysis of Legislation Networks*. Proceedings JURIX 2016.
- N. Sakhaee, S. C. Hendy, M.C. Wilson, G.Zakeri. *Network analysis of New Zealand legislation*. NZ Law Journal 2017.

# Basic setup for discrete time diffusion models

- We focus on learning and beliefs (relevant for psychology, for example). Other applications (e.g. political science, public health) focus on other interpretations (e.g. preferences, disease).

# Basic setup for discrete time diffusion models

- We focus on learning and beliefs (relevant for psychology, for example). Other applications (e.g. political science, public health) focus on other interpretations (e.g. preferences, disease).
- Abstractly, each node has a state (*color*). The state can be discrete or continuous.

# Basic setup for discrete time diffusion models

- We focus on learning and beliefs (relevant for psychology, for example). Other applications (e.g. political science, public health) focus on other interpretations (e.g. preferences, disease).
- Abstractly, each node has a state (*color*). The state can be discrete or continuous.
- At each discrete time step, update a node by a fixed function of the colors of its neighboring nodes.

# Basic setup for discrete time diffusion models

- We focus on learning and beliefs (relevant for psychology, for example). Other applications (e.g. political science, public health) focus on other interpretations (e.g. preferences, disease).
- Abstractly, each node has a state (*color*). The state can be discrete or continuous.
- At each discrete time step, update a node by a fixed function of the colors of its neighboring nodes.
- We study dynamics of the profile of node states. Analytic results are hard for all but the easiest models.

# Fundamental questions

- (equilibrium) Does the process converge in finite time on a given finite $G$? at what rate?

# Fundamental questions

- (equilibrium) Does the process converge in finite time on a given finite $G$? at what rate?
- (unanimity) If it converges, do all nodes have the same color?

# Fundamental questions

- (equilibrium) Does the process converge in finite time on a given finite $G$? at what rate?
- (unanimity) If it converges, do all nodes have the same color?
- (wisdom of crowds) If unanimity is achieved, is it the "correct" color? if not, does the "correct" color win a plurality vote?

# Fundamental questions

- (equilibrium) Does the process converge in finite time on a given finite $G$? at what rate?
- (unanimity) If it converges, do all nodes have the same color?
- (wisdom of crowds) If unanimity is achieved, is it the "correct" color? if not, does the "correct" color win a plurality vote?
- (homophily) Describe the effect on the process of assuming that nodes of same color are more likely to be connected.

# Fundamental questions

- (equilibrium) Does the process converge in finite time on a given finite $G$? at what rate?
- (unanimity) If it converges, do all nodes have the same color?
- (wisdom of crowds) If unanimity is achieved, is it the "correct" color? if not, does the "correct" color win a plurality vote?
- (homophily) Describe the effect on the process of assuming that nodes of same color are more likely to be connected.
- (cascades) When do arbitrary changes to some nodes propagate to a large fraction of the network?

# Some discrete time belief change models

- DeGroot (1974): state is subjective probability, each agent simultaneously averages neighboring states (and own) with some fixed weights. Typically converges via standard Markov chain results.

# Some discrete time belief change models

- DeGroot (1974): state is subjective probability, each agent simultaneously averages neighboring states (and own) with some fixed weights. Typically converges via standard Markov chain results.
- Each agent has a reported $0/1$ belief, and a threshold $0 \leq t \leq 1$, and changes to the other state if fraction at least $t$ of its neighbors have that state.

## Some discrete time belief change models

- DeGroot (1974): state is subjective probability, each agent simultaneously averages neighboring states (and own) with some fixed weights. Typically converges via standard Markov chain results.
- Each agent has a reported $0/1$ belief, and a threshold $0 \leq t \leq 1$, and changes to the other state if fraction at least $t$ of its neighbors have that state.
- The last two can easily oscillate depending on topology and initial coloring. Note that these differ markedly from standard probabilistic contagion models for disease.

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;
- participants iteratively and simultaneously revise answers;

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;
- participants iteratively and simultaneously revise answers;
- feedback to participants is controlled (in particular, open discussion is not allowed);

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;
- participants iteratively and simultaneously revise answers;
- feedback to participants is controlled (in particular, open discussion is not allowed);
- at each iteration, each participant is given statistical feedback about the answers of other participants.

# Iterative distributed jury model

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;
- participants iteratively and simultaneously revise answers;
- feedback to participants is controlled (in particular, open discussion is not allowed);
- at each iteration, each participant is given statistical feedback about the answers of other participants.
- Although this is a very restricted environment, I think it has some relevance to online social networks and political discussion. It is relevant to multiagent intelligent systems.

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.
- We wanted to see whether the model made any sense before investing a lot of resources into analysing it. Unfortunately we invested more resources in experiments ...

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.
- We wanted to see whether the model made any sense before investing a lot of resources into analysing it. Unfortunately we invested more resources in experiments ...
- We performed exploratory laboratory experiments where undergraduates answered “true”, “false” or “don’t know” to objective questions, both famous logical puzzles from psychology and our own experiential questions.

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.
- We wanted to see whether the model made any sense before investing a lot of resources into analysing it. Unfortunately we invested more resources in experiments ...
- We performed exploratory laboratory experiments where undergraduates answered “true”, “false” or “don’t know” to objective questions, both famous logical puzzles from psychology and our own experiential questions.
- Examples:

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.
- We wanted to see whether the model made any sense before investing a lot of resources into analysing it. Unfortunately we invested more resources in experiments ...
- We performed exploratory laboratory experiments where undergraduates answered "true", "false" or "don't know" to objective questions, both famous logical puzzles from psychology and our own experiential questions.
- Examples:
  - (From Frederick's Cognitive Reflection Test) A bat and ball together cost $1.10 and the bat costs $1 more than the ball, so the ball costs $0.10.

# Our work

- We had a new threshold-type model based on the theory of *belief revision* in logic.
- We wanted to see whether the model made any sense before investing a lot of resources into analysing it. Unfortunately we invested more resources in experiments ...
- We performed exploratory laboratory experiments where undergraduates answered "true", "false" or "don't know" to objective questions, both famous logical puzzles from psychology and our own experiential questions.
- Examples:
  - (From Frederick's Cognitive Reflection Test) A bat and ball together cost \$1.10 and the bat costs \$1 more than the ball, so the ball costs \$0.10.
  - The name of the character played by Paul Walker in "The Fast and the Furious" is "Dominic".

# Findings so far

- Subjects answer “don’t know” much less often than expected.

# Findings so far

- Subjects answer “don’t know” much less often than expected.
- Subjects admitting “don’t know” at first iteration learned much more often than those who answer wrongly.

# Findings so far

- Subjects answer “don’t know” much less often than expected.
- Subjects admitting “don’t know” at first iteration learned much more often than those who answer wrongly.
- There is much more apparent social influence than we expected on logical questions.

# Findings so far

- Subjects answer “don’t know” much less often than expected.
- Subjects admitting “don’t know” at first iteration learned much more often than those who answer wrongly.
- There is much more apparent social influence than we expected on logical questions.
- Difficult questions may lead to good social learning, but “tricky” questions (where subjects don’t know they don’t know) lead to really bad social learning.

# Findings so far

- Subjects answer "don't know" much less often than expected.
- Subjects admitting "don't know" at first iteration learned much more often than those who answer wrongly.
- There is much more apparent social influence than we expected on logical questions.
- Difficult questions may lead to good social learning, but "tricky" questions (where subjects don't know they don't know) lead to really bad social learning.
- Promising new model to study: switching probability from "yes" to "no" is proportional to $(p_B^2 - p_W^2)$.

# References


- P. Girard, V. Pavlov, M.C. Wilson. *Networked crowds answer tricky questions poorly*. Preprint 2016.

# References


- P. Girard, V. Pavlov, M.C. Wilson. *Networked crowds answer tricky questions poorly*. Preprint 2016.
- P. Girard, V. Pavlov, M.C. Wilson. *Belief diffusion in social networks*. Preprint 2015.

# References

- P. Girard, V. Pavlov, M.C. Wilson. *Networked crowds answer tricky questions poorly*. Preprint 2016.
- P. Girard, V. Pavlov, M.C. Wilson. *Belief diffusion in social networks*. Preprint 2015.
- Could use help on methodology: how do we falsify a model? how do we get enough data to test a model? what techniques of statistical inference are appropriate?

# Balance in signed networks

- A signed network is an undirected network $G = (V, E)$ together with a map $\sigma : E \to \{\pm 1\}$; write $E_- = \sigma^{-1}(-1)$. A signed graph has a signed adjacency matrix $A$.

# Balance in signed networks

- A signed network is an undirected network $G = (V, E)$ together with a map $\sigma : E \to \{\pm 1\}$; write $E_- = \sigma^{-1}(-1)$. A signed graph has a signed adjacency matrix $A$.
- A cycle is balanced if the product of signs of edges in every cycle is $+1$.

# Balance in signed networks

- A signed network is an undirected network $G = (V, E)$ together with a map $\sigma : E \to \{\pm 1\}$; write $E_- = \sigma^{-1}(-1)$. A signed graph has a signed adjacency matrix $A$.
- A cycle is balanced if the product of signs of edges in every cycle is $+1$.
- $G$ is balanced iff all its cycles are balanced.

# Balance in signed networks

- A signed network is an undirected network $G = (V, E)$ together with a map $\sigma : E \to \{\pm 1\}$; write $E_- = \sigma^{-1}(-1)$. A signed graph has a signed adjacency matrix $A$.
- A cycle is balanced if the product of signs of edges in every cycle is $+1$.
- $G$ is balanced iff all its cycles are balanced.
- Real-world networks are rarely balanced.

# Properties equivalent to balance

- $V = V_0 \cup V_1$ such that $(x, y) \in E_-$ implies $x \in V_i, y \in V_{1-i}$ (polarization).

# Properties equivalent to balance

- $V = V_0 \cup V_1$ such that $(x, y) \in E_-$ implies $x \in V_i, y \in V_{1-i}$ (polarization).
- $(V, E_-)$ is bipartite.

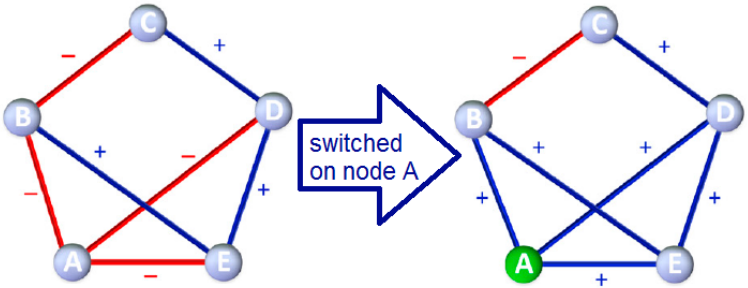
# Properties equivalent to balance

- $V = V_0 \cup V_1$ such that $(x, y) \in E_-$ implies $x \in V_i, y \in V_{1-i}$ (polarization).
- $(V, E_-)$ is bipartite.
- $G$ is switching equivalent to a graph with all positive edges.

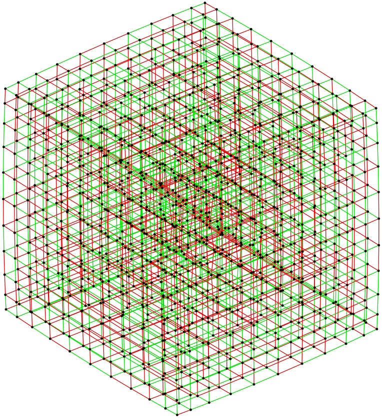
# Properties equivalent to balance

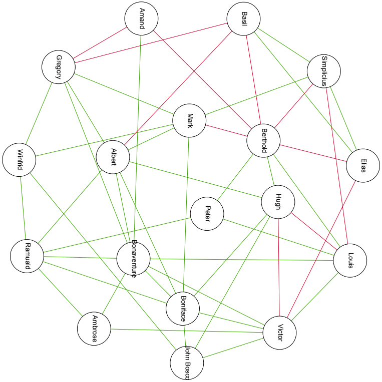
- $V = V_0 \cup V_1$ such that $(x, y) \in E_-$ implies $x \in V_i, y \in V_{1-i}$ (polarization).
- $(V, E_-)$ is bipartite.
- $G$ is switching equivalent to a graph with all positive edges.
- The smallest eigenvalue of the Laplacian $D - A$ (where $D$ is the diagonal matrix of degrees) of $G$ is $0$.

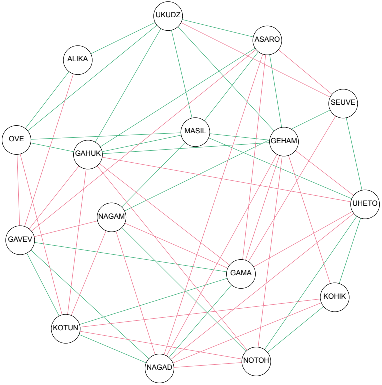
# Switching

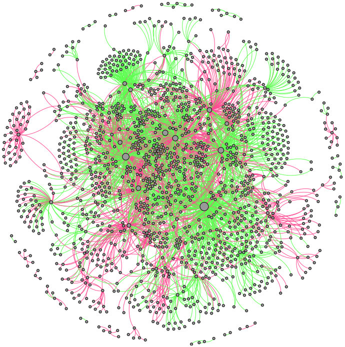


# Examples of real world signed networks

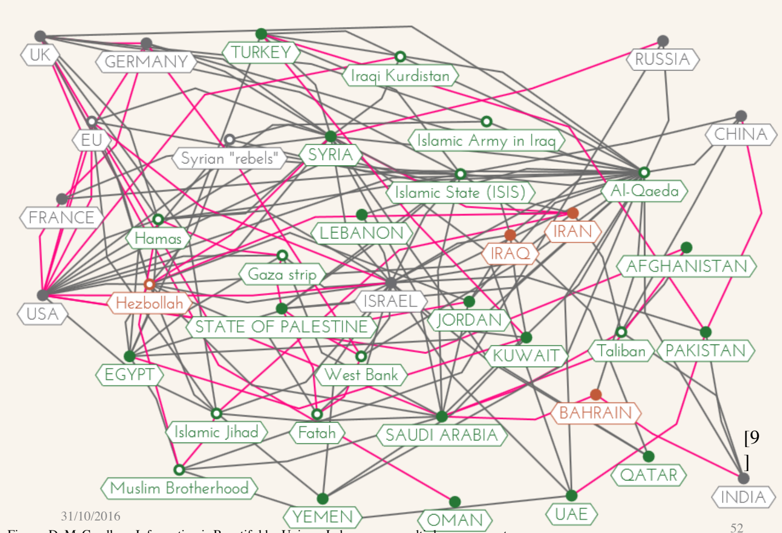





UKUDZ
ASARO
ALIKA
SEUVE
MASIL
OVE
GEHAM
GAHUK
UHETO
NAGAM
GAVEV
GAMA
KOHIK
KOTUN
NAGAD
NOTOH



# Middle East signed network




[9
]

Figure: D. McCandless, Information is Beautiful by Univers Labs, source: multiple news reports

# How to measure partial balance?

- Real networks are not usually balanced, but there are theories that they become more balanced over time.

# How to measure partial balance?

- Real networks are not usually balanced, but there are theories that they become more balanced over time.
- There is no standard measure of partial balance. This has not stopped several papers being written arguing that the above theory is correct or incorrect (!)

# How to measure partial balance?

- Real networks are not usually balanced, but there are theories that they become more balanced over time.
- There is no standard measure of partial balance. This has not stopped several papers being written arguing that the above theory is correct or incorrect (!)
- We reviewed several measures, introduced axioms and desirable properties, and studied them thoroughly on synthetic and real data.

# How to measure partial balance?

- Real networks are not usually balanced, but there are theories that they become more balanced over time.
- There is no standard measure of partial balance. This has not stopped several papers being written arguing that the above theory is correct or incorrect (!)
- We reviewed several measures, introduced axioms and desirable properties, and studied them thoroughly on synthetic and real data.
- One of the best performing measures: the frustration index, a normalization of the minimum number of edges we must flip/delete in order to achieve perfect balance.

# How to measure partial balance?

- Real networks are not usually balanced, but there are theories that they become more balanced over time.
- There is no standard measure of partial balance. This has not stopped several papers being written arguing that the above theory is correct or incorrect (!)
- We reviewed several measures, introduced axioms and desirable properties, and studied them thoroughly on synthetic and real data.
- One of the best performing measures: the frustration index, a normalization of the minimum number of edges we must flip/delete in order to achieve perfect balance.
- We show that well known and commonly used measures such as the fraction of balanced cycles have serious drawbacks.

# Axioms

- A1 $0 \leq \mu(G) \leq 1$.
- A2 $\mu(G) = 1$ if and only if $G$ is balanced.
- A3 If $\mu(G) \leq \mu(H)$, then $\mu(G) \leq \mu(G \oplus H) \leq \mu(H)$.
- A4 $\mu(G^{g(X)}) = \mu(G)$.
- B1 If $\mu(G) \neq 1$, then $\mu(G \oplus C_3^+) > \mu(G)$.
- B2 If $\mu(G) \neq 0$, then $\mu(G \oplus C_3^-) < \mu(G)$
- B3 If $e \in E^*$, then $\mu(G \ominus e) \geq \mu(G)$.
- B4 If $\mu(G) \neq 0$ and $\mu(G \ominus E^* \oplus e) \neq 1$, then $\mu(G \oplus e) \leq \mu(G)$.

# Axiomatic behavior of measures

| | $D(G)$ | $C(G)$ | $W(G)$ | $D_k(G)$ | $A(G)$ | $F(G)$ |
|---|---|---|---|---|---|---|
| A1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| A2 | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| A3 | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| A4 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| B1 | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| B2 | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| B3 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| B4 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |

# Balance in minimally and maximally unbalanced $K_n$

| $\mu(G)$ | $\mu(G_{\min})$ | $\mu(G_{\max})$ |
|---|---|---|
| $D(G)$ | $\sim 1 - 2/n$ | $\sim \frac{1}{2} + (-1)^n e^{-2}$ |
| $C(G)$ | $\sim 1 - 1/n$ | $\sim \frac{1}{2} - \frac{3n \log n}{2^n}$ |
| $D_k(G)$ | $1 - 2k/n(n-1)$ | $0, 1$ |
| $W(G)$ | $\sim 1 - 2/n$ | $\sim \frac{1+e^{2-2n}}{2}$ |
| $A(G)$ | $\sim 1 - 4/n^2$ | $0$ |
| $F(G)$ | $1 - 4/n(n-1)$ | $\frac{1}{n}, \frac{1}{n-1}$ |

# How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.

# How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.
- However we still need to compute it. We started with a basic integer programming model and now have 3 models:

## How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.
- However we still need to compute it. We started with a basic integer programming model and now have 3 models:
  - using data reduction (preprocessing)

# How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.
- However we still need to compute it. We started with a basic integer programming model and now have 3 models:
  - using data reduction (preprocessing)
  - reformulating our original nonlinear model as a linear one

## How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.
- However we still need to compute it. We started with a basic integer programming model and now have 3 models:
  - using data reduction (preprocessing)
  - reformulating our original nonlinear model as a linear one
  - using the structure of the problem to create nonobvious constraints

# How to compute the frustration index?

- It is known that computing the frustration index is NP-hard in general (by reduction from MAX-CUT). It is equivalent to minimizing edges between vertices of the same color, over all possible vertex colorings.
- However we still need to compute it. We started with a basic integer programming model and now have 3 models:
  - using data reduction (preprocessing)
  - reformulating our original nonlinear model as a linear one
  - using the structure of the problem to create nonobvious constraints
  - using IP techniques ("lazy cuts")

# XOR model

$$
\begin{aligned}
\min_{x_i : i \in V, f_{ij} : (i,j) \in E} Z = & \sum_{(i,j) \in E} f_{ij} \\
\text{s.t.} \quad & f_{ij} \geq x_i - x_j \quad \forall (i,j) \in E^+ \\
& f_{ij} \geq x_j - x_i \quad \forall (i,j) \in E^+ \\
& f_{ij} \geq x_i + x_j - 1 \quad \forall (i,j) \in E^- \\
& f_{ij} \geq 1 - x_i - x_j \quad \forall (i,j) \in E^- \\
& x_i \in \{0,1\} \quad \forall i \in V \\
& f_{ij} \in \{0,1\} \quad \forall (i,j) \in E
\end{aligned} \tag{1}
$$

# Some additional constraints

Feasible:

$$f_{ij} + f_{ik} + f_{jk} \geq 1 \quad \forall (i,j,k) \in T^-$$

Optimal:

$$\sum_{j:(i,j)\in E \text{ or } (j,i)\in E} f_{ij} \leq (d_i/2) \quad \forall i \in V$$

# Results

- The current implementations are at least 10 times faster than the original and allow computation in networks with thousands of nodes and edges.

# Results

- The current implementations are at least 10 times faster than the original and allow computation in networks with thousands of nodes and edges.
- They are the best we know of by quite some distance (several orders of magnitude faster on test problems).

# Results

- The current implementations are at least 10 times faster than the original and allow computation in networks with thousands of nodes and edges.
- They are the best we know of by quite some distance (several orders of magnitude faster on test problems).
- We show that many previously computed results are incorrect.

# Sample results

| | Graph | D2007 | H2010 | I2010 | XOR |
|---|---|---|---|---|---|
| Quality | EGFR | [196, 219] | 210 | [186, 193] | 193 |
| | macrophage | [218,383] | 374 | [302, 332] | 332 |
| | yeast | [0, 43] | 41 | 41 | 41 |
| | E.coli | [0, 385] | fail | [365, 371] | 371 |
| Time | EGFR | 420 s | 6480 s | >60 s | 0.28 s |
| | macrophage | 2640 s | 60 s | >60 s | 0.56 s |
| | yeast | 4620 s | 60 s | >60 s | 0.13 s |
| | E.coli | - | fail | >60 s | 2.21 s |

# Are real-world networks more balanced than random ones?

- Using the normalized frustration index shows that many are.

# Are real-world networks more balanced than random ones?

- Using the normalized frustration index shows that many are.
- Social and political networks (e.g. New Guinea highland tribes, social relations in a monastery, Senate bill co-sponsorship) and some biological networks (e.g. gene regulatory networks) are much more balanced than expected by chance.

# Are real-world networks more balanced than random ones?

- Using the normalized frustration index shows that many are.
- Social and political networks (e.g. New Guinea highland tribes, social relations in a monastery, Senate bill co-sponsorship) and some biological networks (e.g. gene regulatory networks) are much more balanced than expected by chance.
- However certain biological networks are much less balanced than expected.

# Are real-world networks more balanced than random ones?

- Using the normalized frustration index shows that many are.
- Social and political networks (e.g. New Guinea highland tribes, social relations in a monastery, Senate bill co-sponsorship) and some biological networks (e.g. gene regulatory networks) are much more balanced than expected by chance.
- However certain biological networks are much less balanced than expected.
- International alliances network seems to become (slowly) more balanced over time.

# References

- S. Aref, M.C. Wilson. *Measuring Partial Balance in Signed Networks*. Accepted Journal of Complex Networks 2017.

# References


- S. Aref, M.C. Wilson. *Measuring Partial Balance in Signed Networks*. Accepted Journal of Complex Networks 2017.
- S. Aref, A. J. Mason, M.C. Wilson. *Computing the Line Index of Balance Using Integer Programming Optimisation*. Accepted Graphs' Optimization Problems and Their Computational Complexities (G. Gutin 60th birthday special volume), Springer 2017.

# References


- S. Aref, M.C. Wilson. *Measuring Partial Balance in Signed Networks*. Accepted Journal of Complex Networks 2017.
- S. Aref, A. J. Mason, M.C. Wilson. *Computing the Line Index of Balance Using Integer Programming Optimisation*. Accepted Graphs' Optimization Problems and Their Computational Complexities (G. Gutin 60th birthday special volume), Springer 2017.
- S. Aref, A.J. Mason, M.C.Wilson *An exact method for computing the frustration index in signed networks using binary programming*. Submitted Journal of Combinatorial Optimization 2017.