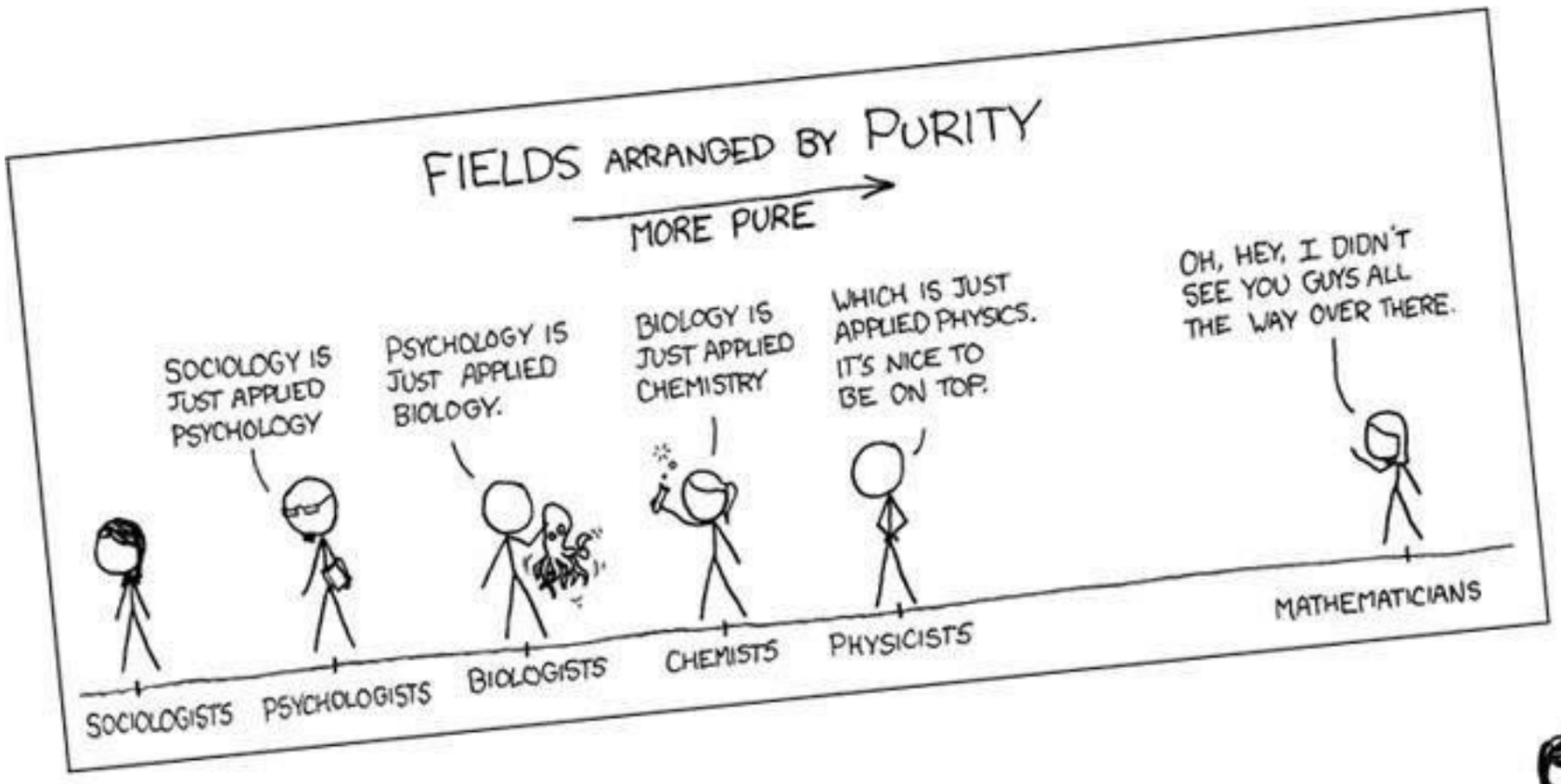
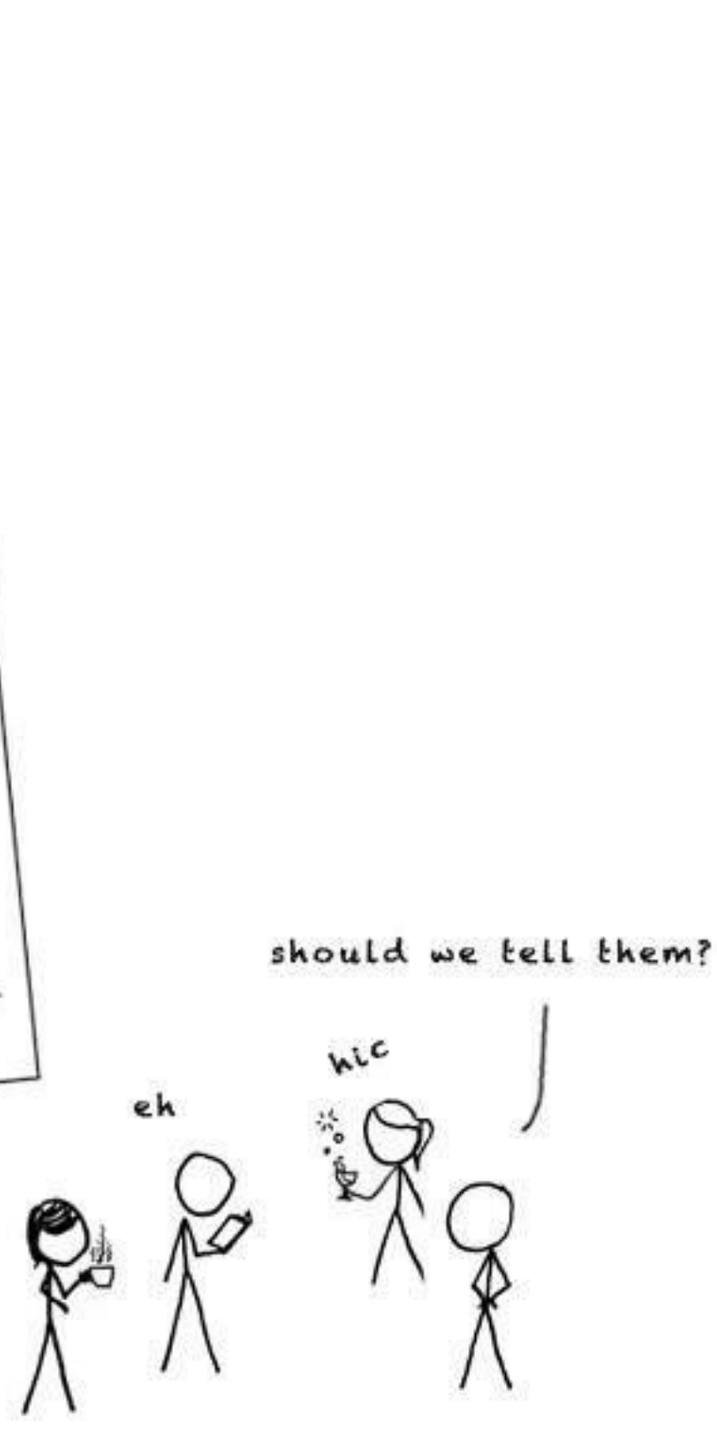
Computational social science: opportunities and risks

Dr Giuseppe A. Veltri

UNIVERSITY OF LEICESTER





philosophers

http://xkcd.com/435/ "fixed" by @rocza

Data revolution?

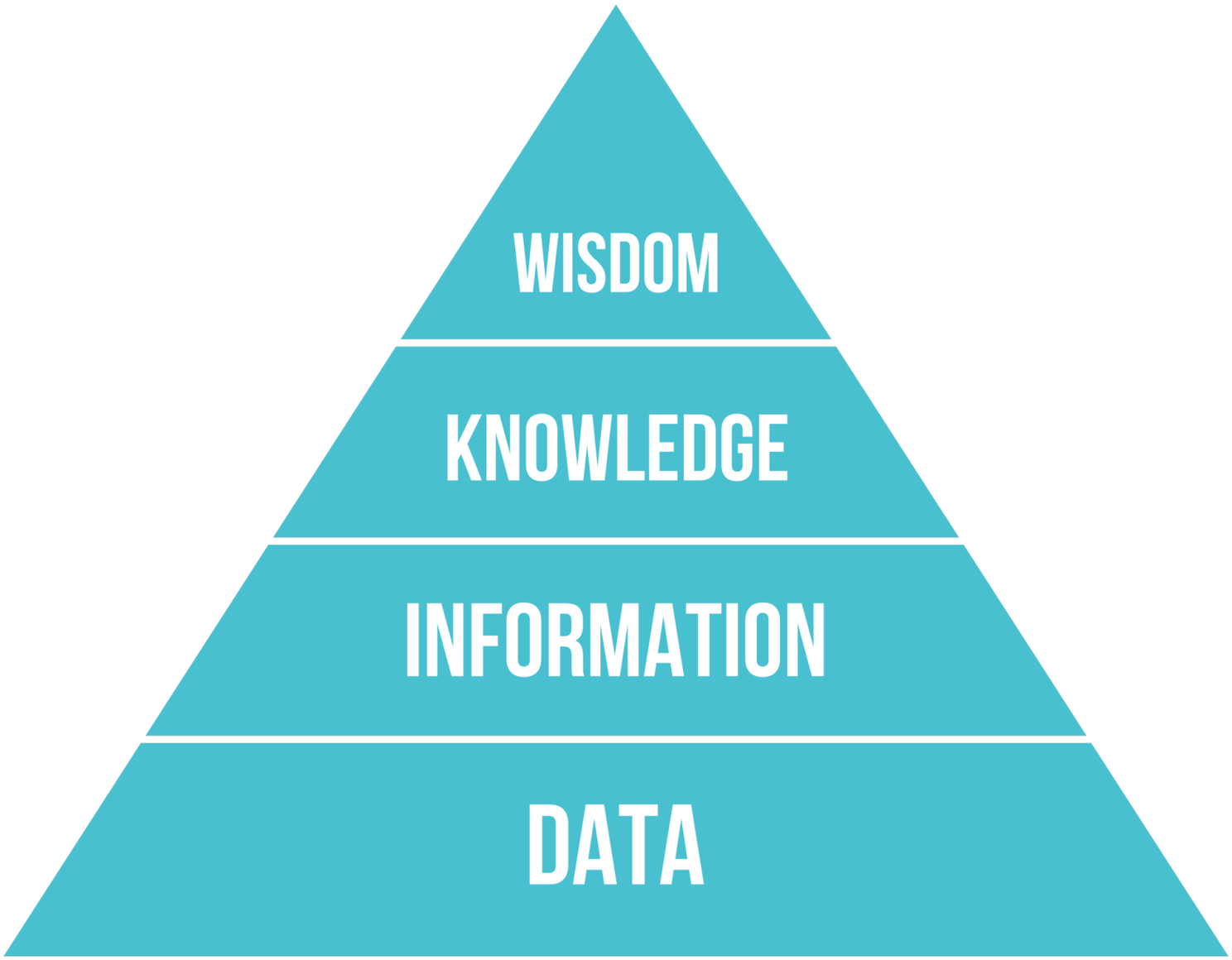
- Revolutions in science have often been preceded by revolutions in measurement

- The availability of big data and data infrastructures, coupled with new analytical tools, challenges established epistemologies

- New answers to old (research) questions or simply new questions?

- True interdisciplinary opportunity

There is nothing more practical than good theory (K. Lewin) but there is a lot of ‘cheap’ theory out there


WISDOM
KNOWLEDGE
INFORMATION
DATA

- The data revolution offers the possibility of shifting:
  - from *data-scarce* to *data-rich* studies of societies;
  - from *static snapshots* to *dynamic unfoldings*; from *coarse aggregations* to *high resolutions*; from relatively *simple models* to more complex, *sophisticated simulations*.

Computational social science

- The information-processing paradigm of CSS has dual aspects: substantive and methodological. *From the substantive point of view*, this means that CSS uses information-processing as a key ingredient for explaining and understanding how society and human beings within it operate to produce emergent complex systems. As a consequence, this also means that social complexity cannot be understood without highlighting human and social processing of information as a fundamental phenomenon.

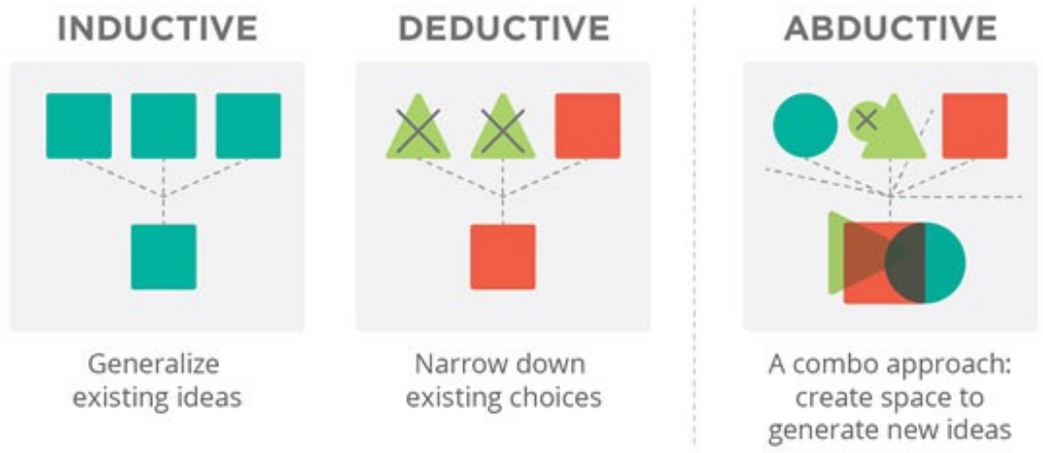
- From a *methodological point of view*, the information-processing paradigm points toward computing as a fundamental instrumental approach for modelling and understanding social complexity. This does not mean that other approaches, such as historical, statistical, or mathematical, become irrelevant.

New epistemology

- Data driven science combines abductive, inductive and deductive approaches.

- It differs from traditional deductive design in that it seeks to generate hypotheses and insights ‘born from the data’ rather than ‘born from the theory’.

- In other words, it seeks to incorporate a mode of induction into the research design, though explanation through induction is not the intended end point.



Knowledge discovery techniques

- Instead, it forms a new mode of hypotheses generation before a deductive approach is employed.

- The epistemological strategy adopted within data driven science is to use guided knowledge discovery techniques to identify potential questions (hypotheses) worth of further examination and testing.

Network science

- Network science is an academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors represented by nodes (or vertices) and the connections between the elements or actors as links (or edges).

- In the context of social sciences, it has been very difficult to collect *relational data*, data about people's interactions. In the recent past, there were only two ways: direct *observations*; *asking people using surveys*. Both are extremely limited.

- The abundance of relational data online has changed this. This is way a lot of social scientists are so eager to use Twitter and Facebook data.

Organic data

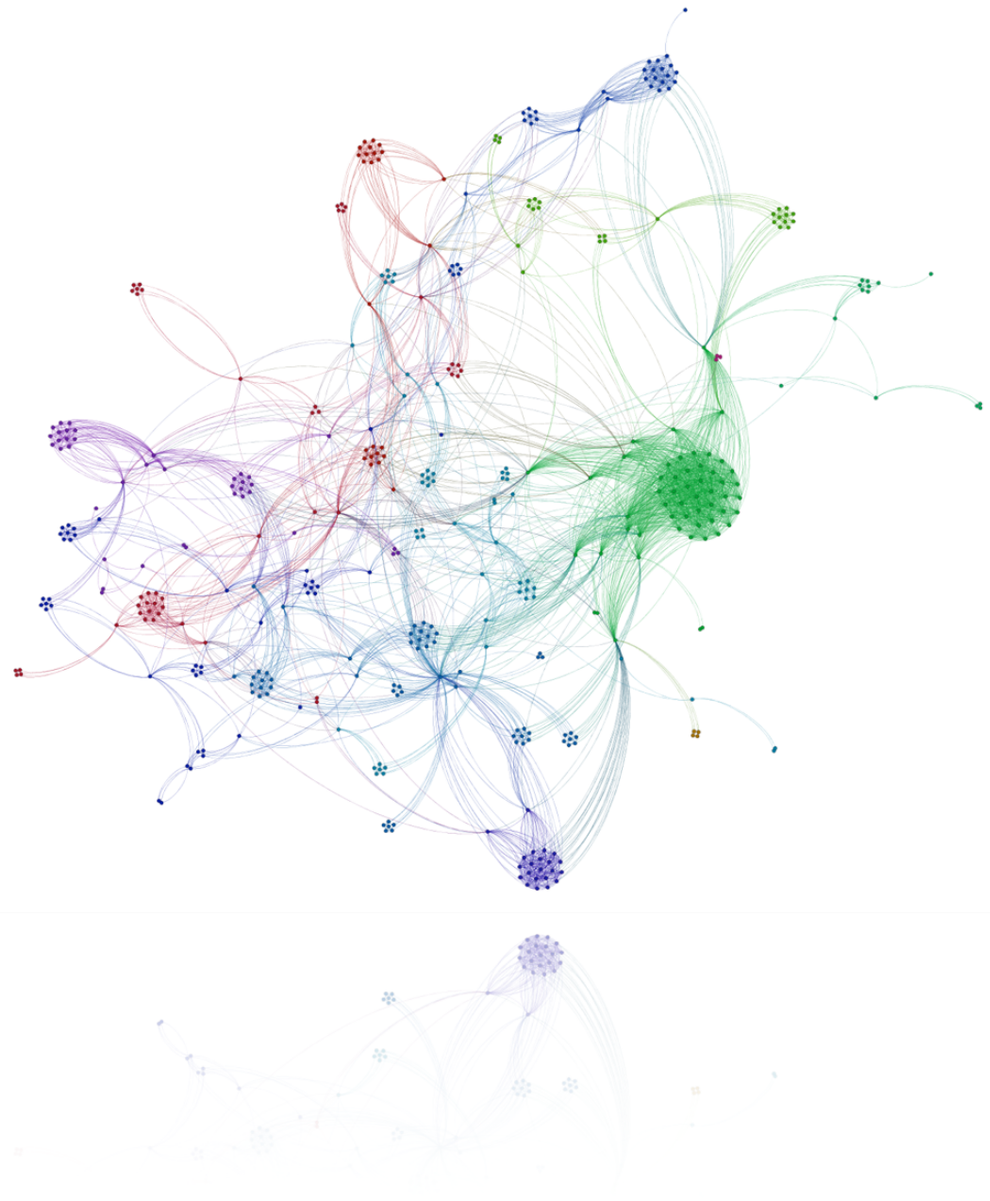
- We're entering a world where data will be the cheapest commodity around, simply because the society has created systems that automatically track transactions of all sorts.
    - For example, internet search engines build data sets with every entry, Twitter generates tweet data continuously, traffic cameras digitally count cars, scanners record purchases, Internet sites capture and store mouse clicks.
- Collectively, the society is assembling data on massive amounts of its behaviours.
- Indeed, if you think of these processes as an ecosystem, it is self-measuring in increasingly broad scope. Indeed, we might label these data as **"organic",** a now-natural feature of this ecosystem.

Designed & Organic data

- Collectively, the society is assembling data on massive amounts of its behaviours.

- We can label these data as ‘**organic**’, a now-natural feature of this ecosystem. Information is produced from data by uses.

- This is in contrast with ‘**designed**‘ data, those that are collected when you design experiment, a questionnaire, a focus group, etc. and to not exist until are collected.



Long data

- Perhaps, the most annoying problem of your research endeavours
- Coping strategies for lack of long data
  - Cross-sectional illusion of control
  - Ignoring decay
  - Processes vs structures

Risks

Ethical risks:
covert research
privacy
transparency
etc.

Simplification of human agency

- E.g. Is a Tweet someone's opinion?
- Does online behaviour mirror offline one?

Correlational studies

- Finding a lot of patterns, for example correlations are a good starting point but not that interesting from the point of view of many social scientists.

- The problem here is a clash of ‘cultures of modelling’ between how we model in the social sciences and how

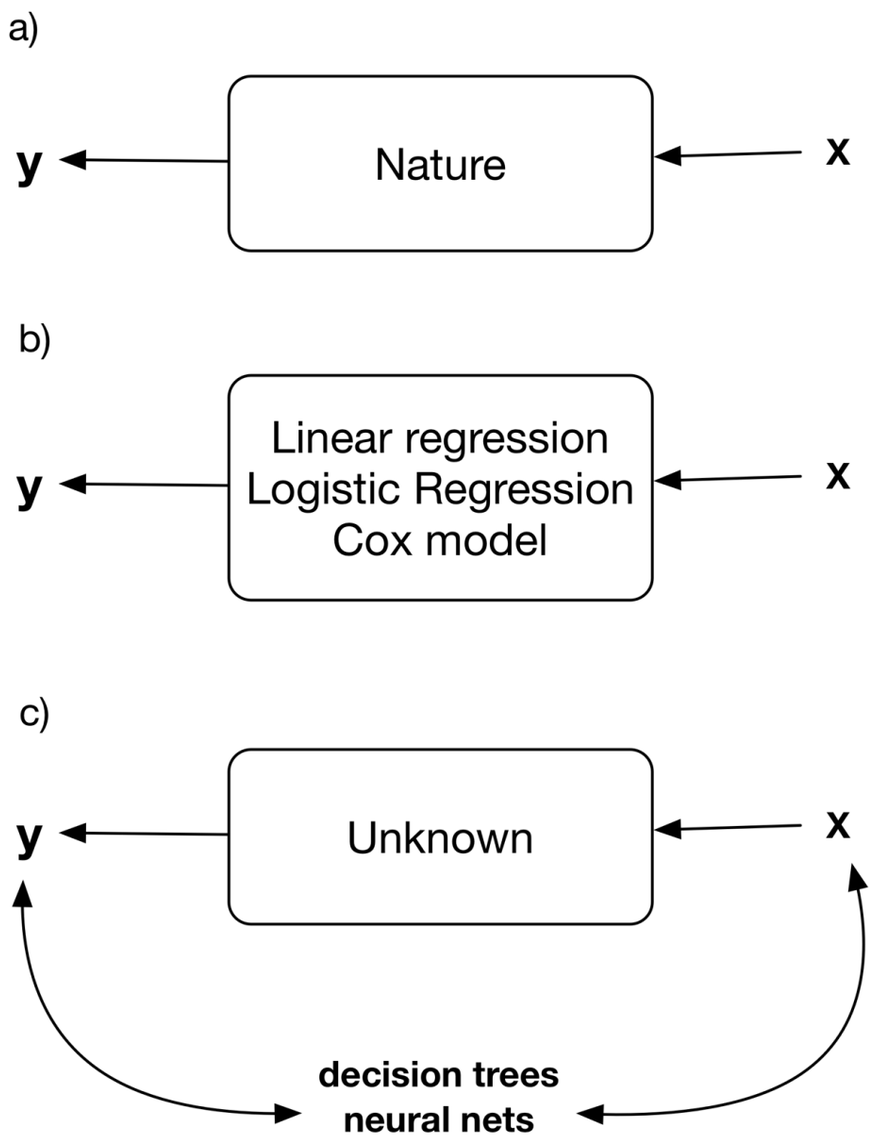
Part 2

The two culture of modelling

- The role of big data and its impact on social science research needs to be addressed in the context of the ‘computational and algorithmic turn’ that is increasingly affecting social science research methods. In order to fully appreciate such a turn, we can contrast the difference between the ‘two cultures of modelling’ (Gentle et al. 2012; Breiman 2001).

- The first is the ‘data modelling’ culture in which the analysis starts by assuming a stochastic data model for the inside of the black box of Figure 1A and therefore resulting in Figure 1B.

- The ‘algorithmic modelling’ considers the inside of the box as complex and unknown. Such an approach is to find an algorithm that operates on x to predict the responses y.



- Borrowing from Breiman (2001), the data modelling approach is about evaluating the values of parameters from the data and after that the model is used for either information or prediction (Figure 1B). In the algorithmic modelling approach, there is a shift from data models to the properties of algorithms.

Classification & regression trees

- Classification and regression trees are based on a purely data-driven paradigm. Without referring to a concrete statistical model, they search recursively for groups of observations with similar values of the response variable by building a tree structure.

- If the response is categorical, one refers to classification trees; if the response is continuous, one refers to regression trees.

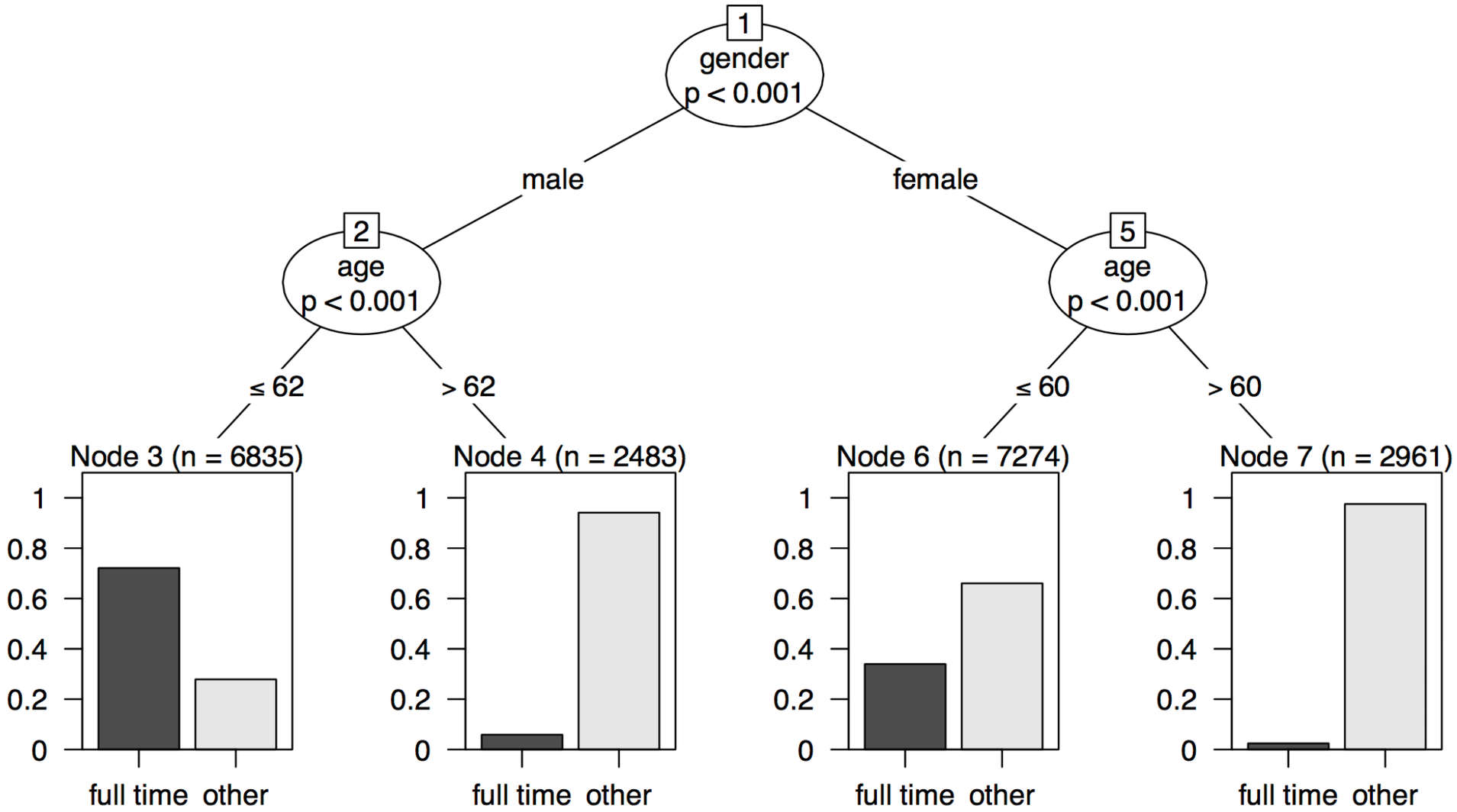


Figure 1: Classification tree: Assessing different frequencies of full-time jobs in Germany (SOEP 2008). The resulting tree-structure shows varying participation rates in full-time labor in three splits according to the covariates gender and age.

```
> library(“party’)
> ct_obj <- ctree(job_time ~ gender + age,
>           control = ctree_control(minsplit = 50),
>           data = data_empl
>
> ct_obj

Conditional inference tree with 4 terminal nodes
     Response:  job_time
     Inputs:  gender, age
     Number of observations:  19553
     1) gender == {male}; criterion = 1, statistic = 1910.231
       2) age <= 62; criterion = 1, statistic = 1397.736
         3)*  weights = 6835
       2) age > 62
         4)*  weights = 2483
     1) gender == {female}
       5) age <= 60; criterion = 1, statistic = 530.524
         6)*  weights = 7274
       5) age > 60
         7)*  weights = 2961
```

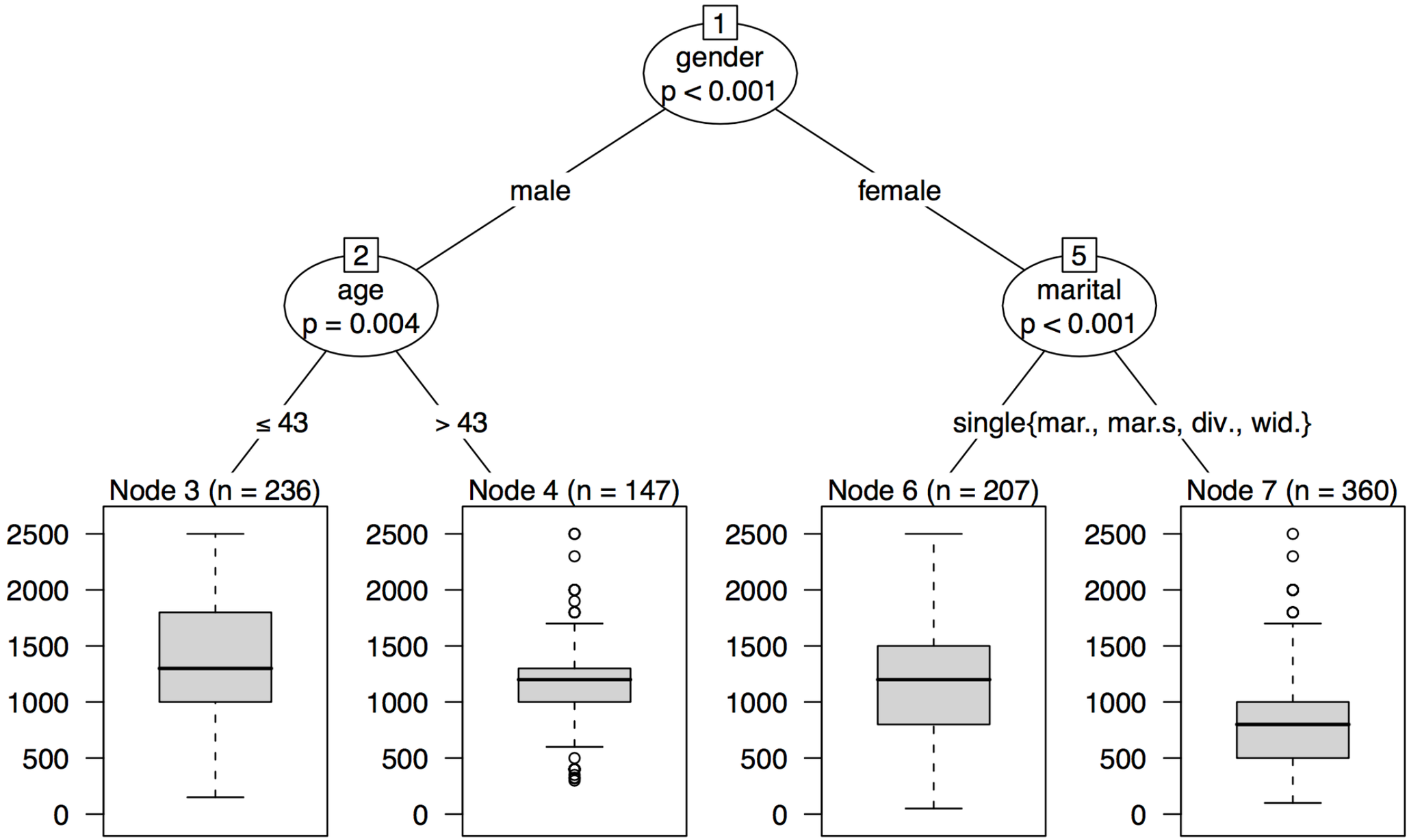


Figure 2: Regression tree: Assessing different requested incomes of unemployed respondents (SOEP 2008). Three different levels are obtained in groups related to gender, age and marital status.

```
> rt_obj <- ctree(take_job ~ gender + age + nation + marital,
>           control = ctree_control(minsplit = 10), data =dat_unempl)

>
> rt_obj

Conditional inference tree with 4 terminal nodes
     Response:  take_job
     Inputs:  gender, age, nation, marital
     Number of observations:  950
     1) gender == {male}; criterion = 1, statistic = 115.915
       2) age <= 43; criterion = 0.988, statistic = 8.841
         3)*  weights = 236
       2) age > 43
         4)*  weights = 147
     1) gender == {female}
       5) marital == {single}; criterion = 1, statistic = 49.76
         6)*  weights = 207
       5) marital == {mar., mar.s, div., wid.}
         7)*  weights = 360
```

Model based recursive portioning

- The method of model-based recursive partitioning forms an advancement of classification and regression trees, which are widely used in life sciences.

- Model-based recursive partitioning (Zeileis et al. 2008) represents a synthesis of a theory- based approach and a data-driven set of constraints to the theory validation and further development.

- In extreme synthesis, this approach works through the following steps.

    1. First, a parametric model is defined to express a theory-driven set of hypotheses (e.g. a linear regression).
    2. Second, this model is evaluated to the model-based recursive partitioning algorithm that checks whether other important covariates have been omitted that would alter the parameters of the initial model

- The same tree-structure of a regression, or classification tree, is produced.

- This time, rather than splitting for different patterns of the response variable, the model-based recursive partitioning *finds different patterns of associations between the response variable and other covariates that have been pre-specified in the parametric model*.

- In other words, it creates different versions of β the model in terms of estimation, depending on different important values of covariates

requestedincome(`jobvar`)=$\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \varepsilon$.

- Here, a linear regression model is investigated. Thus, the linear model explains the dependent variable `jobvar` through the independent variables `age + age2` and a *u*-shaped relationship between the requested income and the predictor variable age is assumed. I

```
> mob_obj <- mob(jobvar ~ age + I(age^2) | gender + nation + marital,
>    control = mob_control(minsplit = 30), data = dat_job,
>    model = linearModel)


> temp <- coef(mob_obj)
> colnames(temp) <- c("Intercept", "age", "age sq.")
> printCoefmat(temp)

Intercept       age age sq.

2   998.916  22.613 -0.3640
4   748.667  11.673 -0.1204
5  1229.166 -17.144  0.1808
```

- In *R* code, the quadratic term for age is generated if `I(age^2)` is included in the formula (arithmetic operations have a different meaning in the formula context and the interpretation is inhibited using `I()`)

- After a vertical bar, potential spitting variables are included, such as `gender + nation + marital` in the example. Here the control argument is `control = mob_control(minsplit = 30, verbose=TRUE)`, allowing, e.g., to specify minimum splitting node sample sizes or to print test statistics during the computation process via `verbose=TRUE`

node	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
2	1014.5837	22.5446	-0.3618
4	1212.6621	-0.0236	-0.0871
5	1390.9708	-25.3983	0.2737

Clearly, the initial model is insufficient to explain such relationship without taking some of these covariates into consideration.

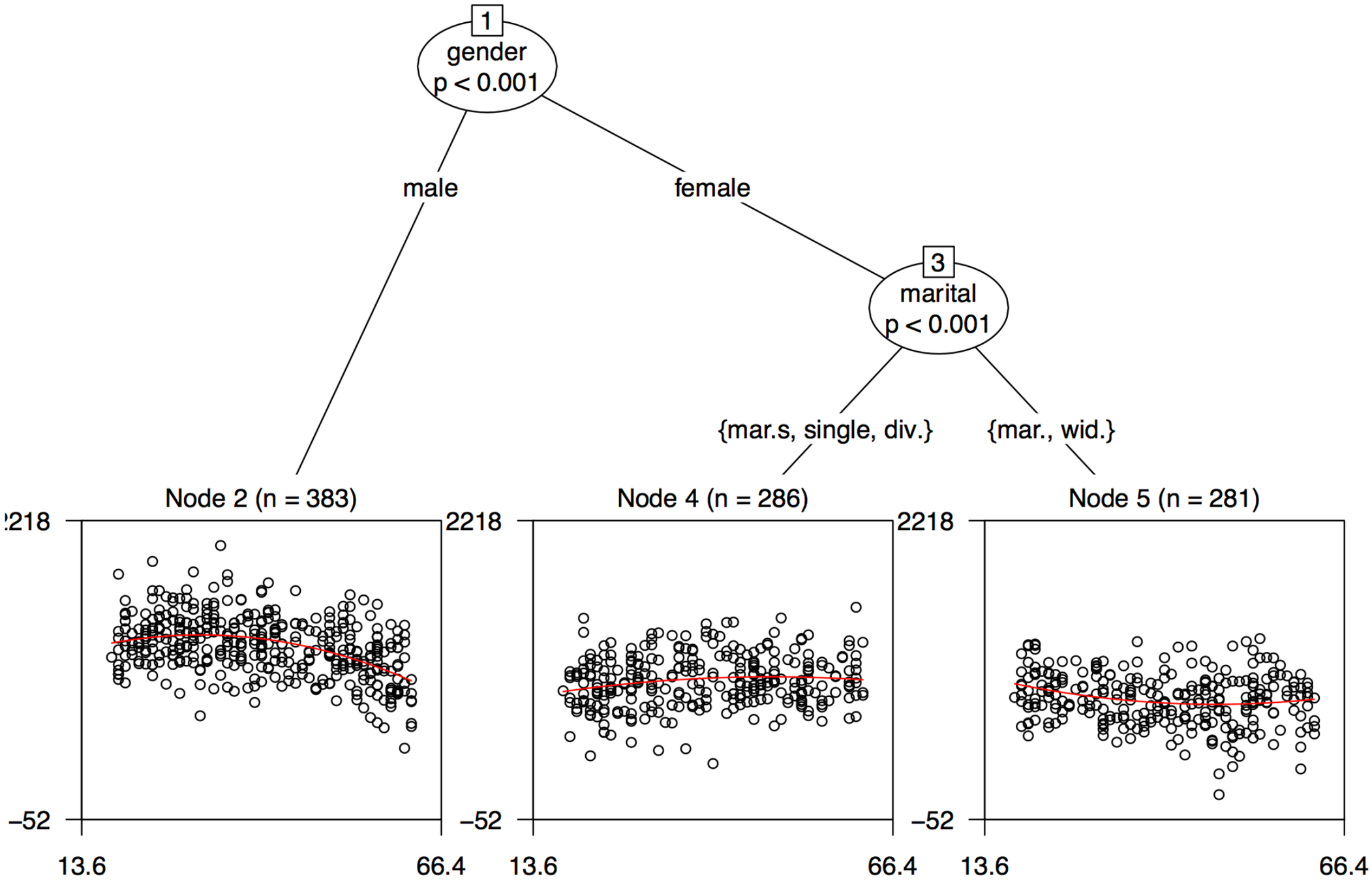


Figure 3.3: Model-based recursive partitioning: Simulated relationship between age and requested income.

- What is the advantage of having such information? The answer to this question refers to the initial distinction that was introduced about the two cultures of modelling.

- In the predominant (in social sciences) data modelling culture, the comparison between different models has always been difficult and a problematic point.

- The hybrid approach of model-based recursive partitioning modelling can help revise models that work for the full dataset and that do not neglect such information imposing on models, as 'global' strait jackets.

- Moreover, if the researcher in question values the working rule of *Ockham's Razor* (that a model should be no more complex than necessary but needs to be complex enough to describe the empirical data), model-based recursive partitioning can be used for evaluating different models.

- One more useful item of information, generated by this approach, is that the model-based recursive method allows the identification of particular segments of the sample under examination that might be worth further investigation.

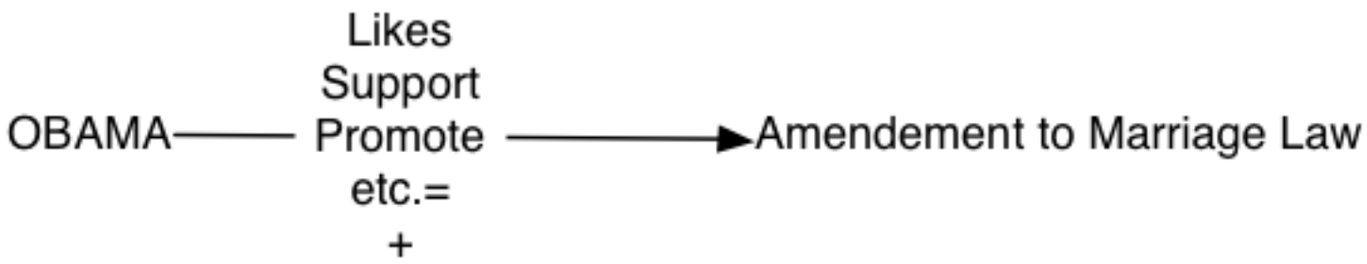
- **Traditionally, we could not use this partitioning of data because we had small datasets.**

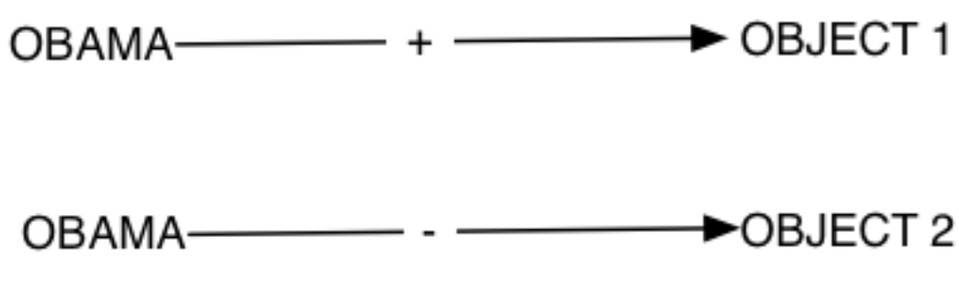
- **Therefore, model comparisons and local models were impossible to detect**

Thank you!

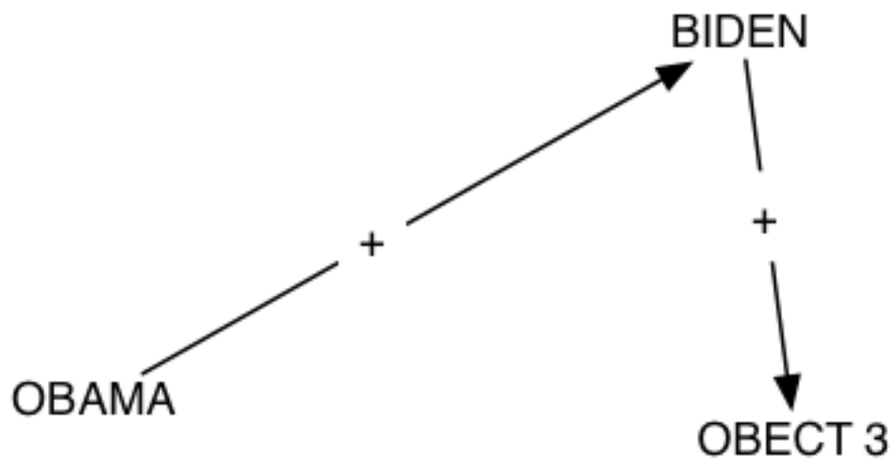
Email: g.a.veltri@le.ac.uk

Massive amounts of text and computations

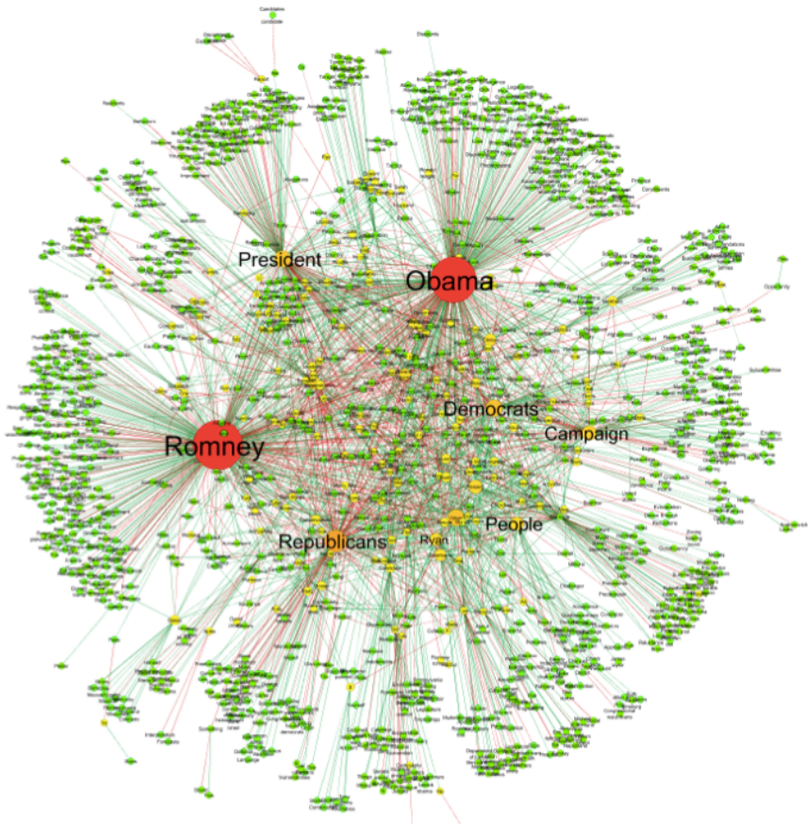

OBAMA
Likes
Support
Promote
etc.=
+
Amendement to Marriage Law

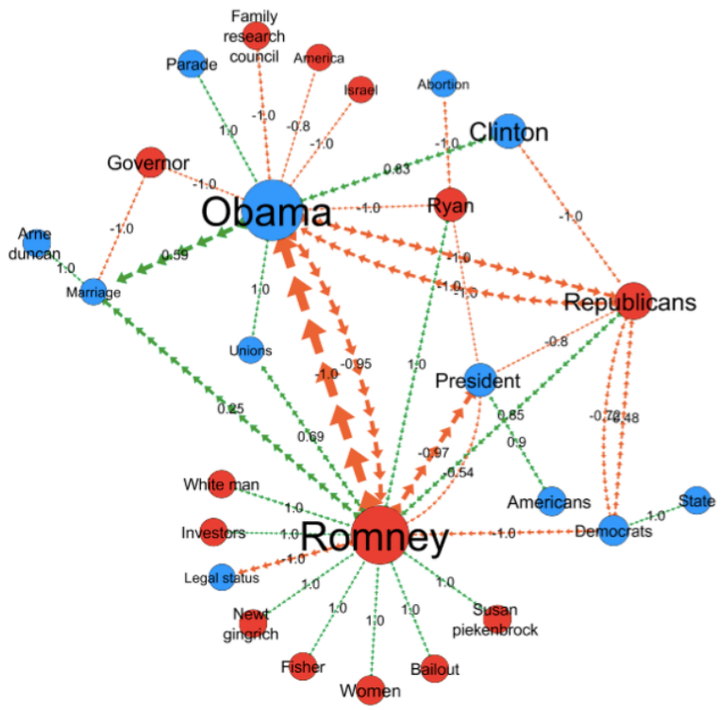

OBAMA
+
OBJECT 1

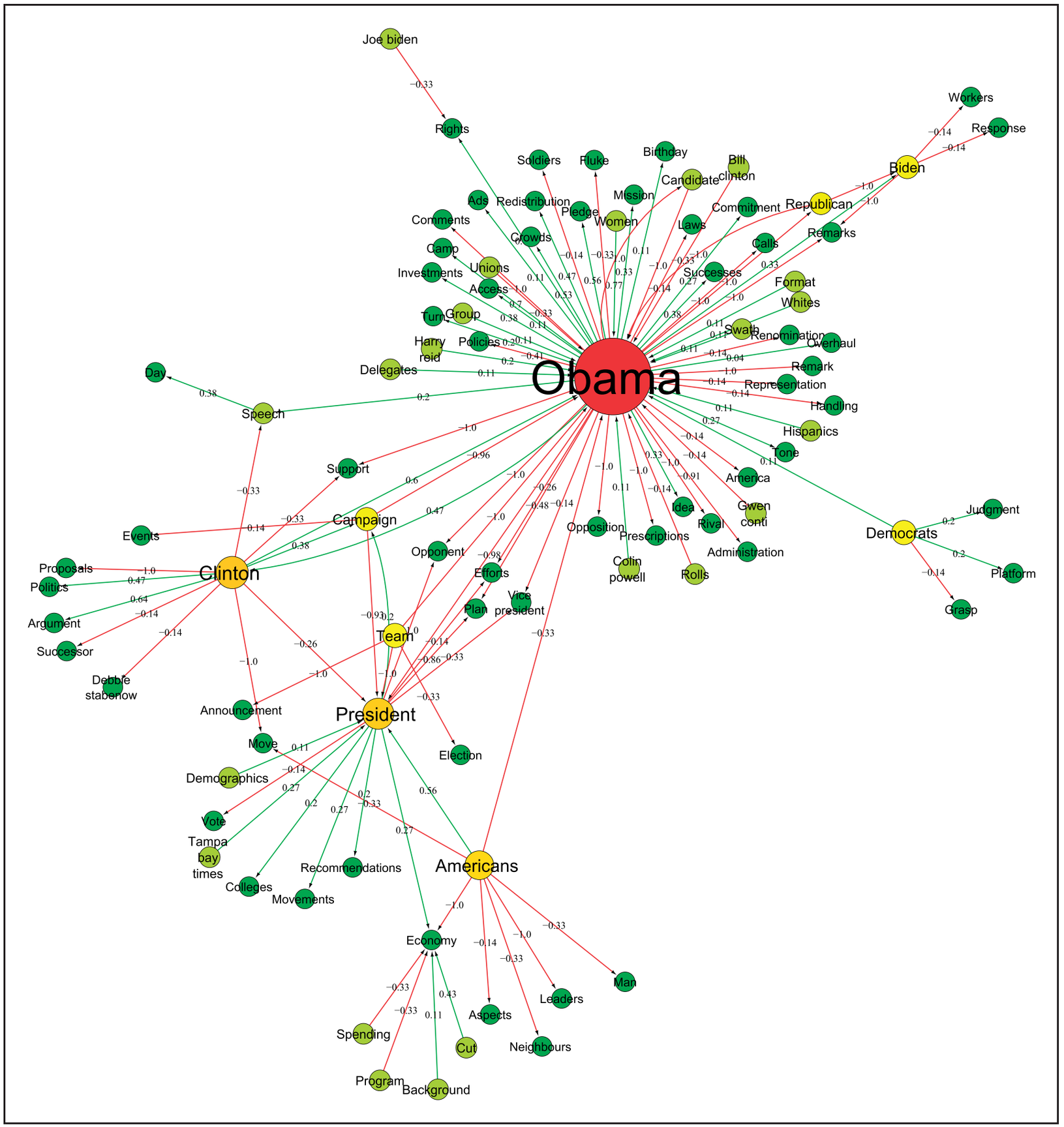
OBAMA
-
OBJECT 2

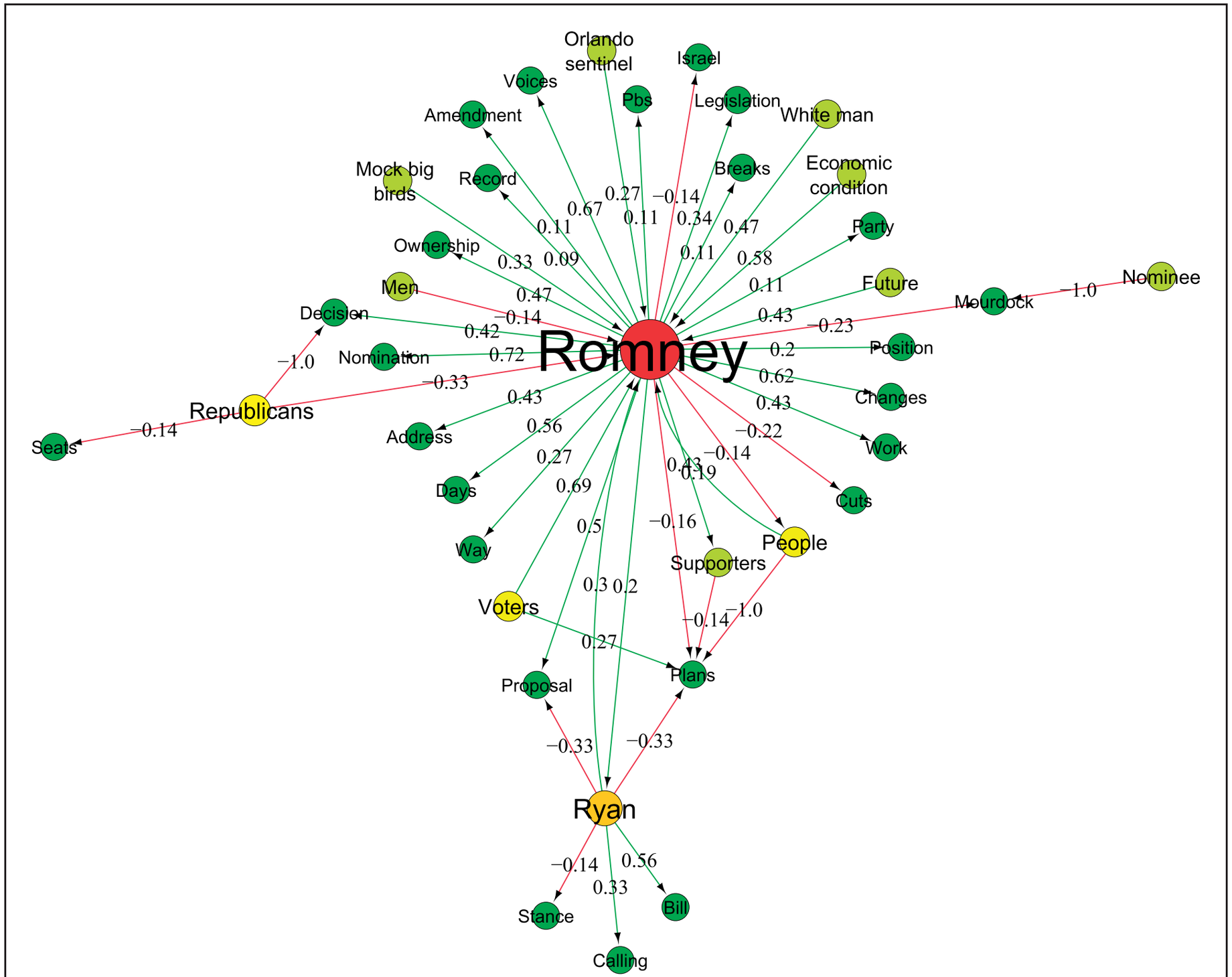

BIDEN
+
+
OBAMA
OBECT 3

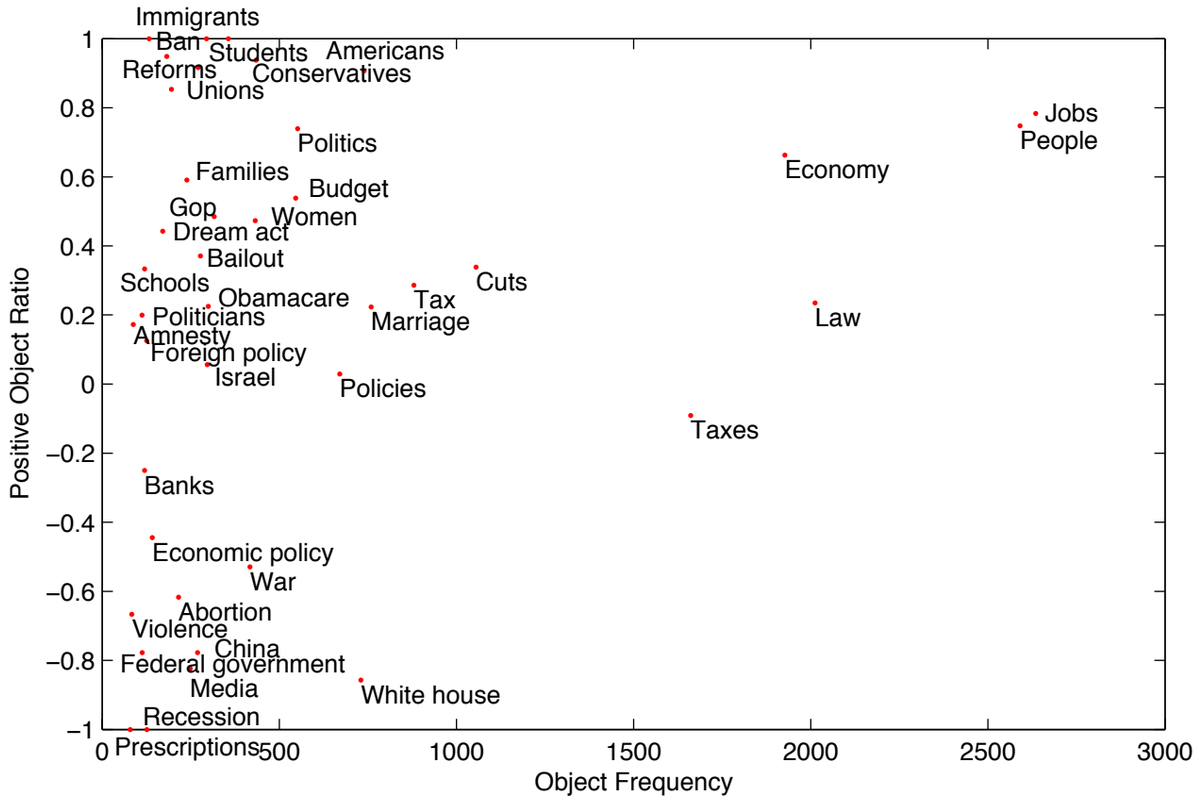
Mediascapes of last US presidential election using advanced text mining





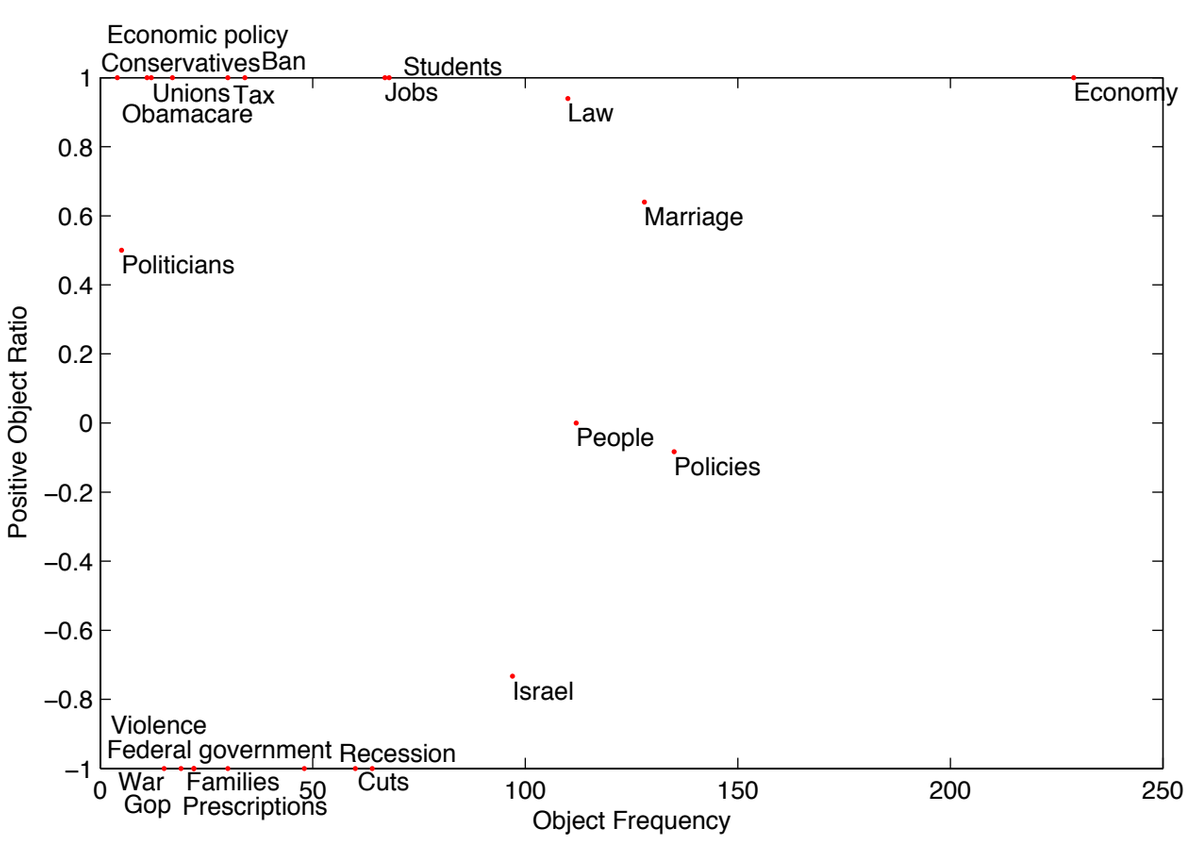

Joe biden
Rights
Soldiers
Fluke
Birthday
Bill clinton
Candidate
Mission
Workers
Response
Biden
Republican
Commitment
Pledge
Women
Laws
Remarks
Ads
Redistribution
Comments
Crowds
Camp
Calls
Unions
Investments
Access
Successes
Format
Whites
Turn
Group
Swath
Renomination
Overhaul
Harry reid
Policies
Delegates
Obama
Remark
Representation
Day
Handling
Speech
Hispanics
Tone
Support
America
Campaign
Idea
Gwen conti
Judgment
Events
Opposition
Rival
Democrats
Opponent
Prescriptions
Administration
Proposals
Clinton
Efforts
Colin powell
Rolls
Platform
Politics
Vice president
Plan
Argument
Team
Grasp
Successor
Debbie stabenow
Announcement
President
Move
Election
Demographics
Vote
Tampa bay times
Recommendations
Americans
Colleges
Movements
Economy
Leaders
Aspects
Spending
Cut
Neighbours
Program
Background

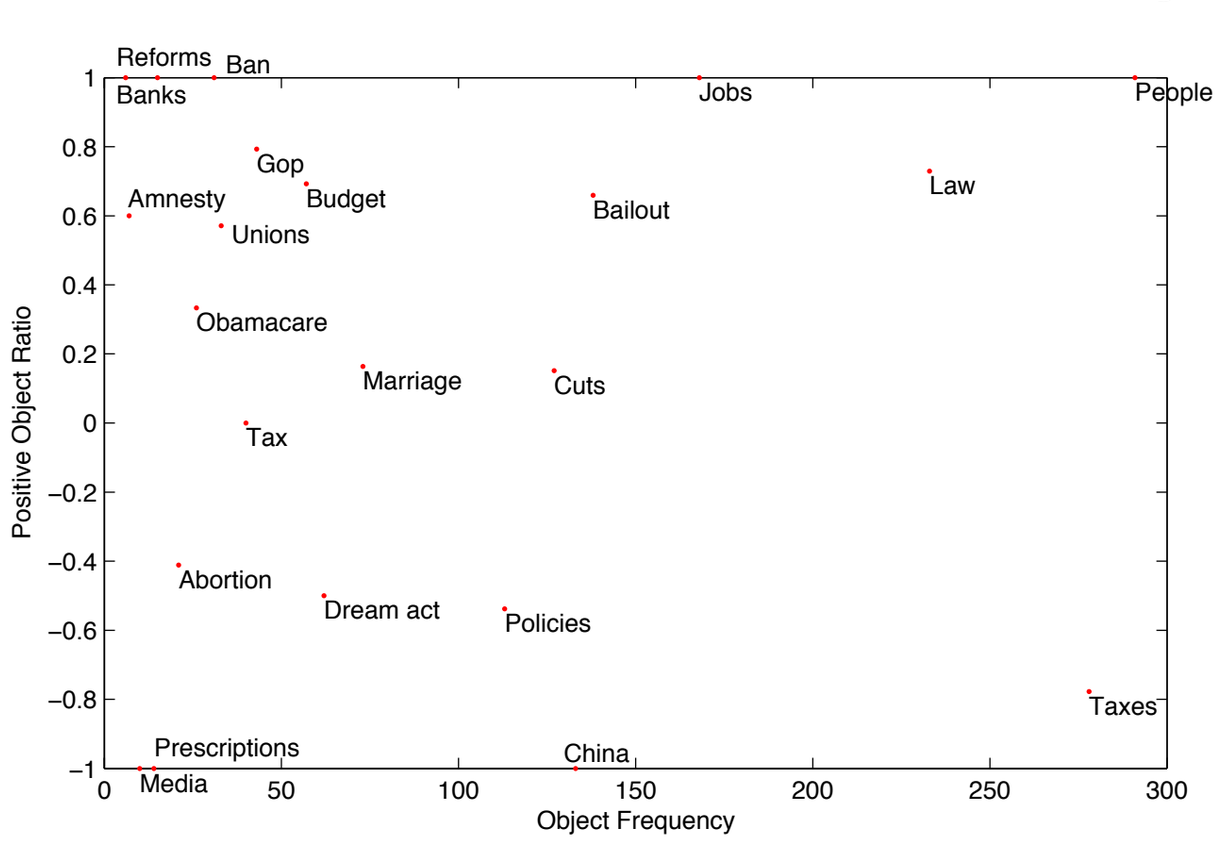

Romney
Republicans
Ryan
Orlando sentinel
Israel
Voices
Pbs
Legislation
Amendment
White man
Mock big birds
Record
Breaks
Economic condition
Party
Ownership
Men
Future
Nominee
Mourdock
Decision
Nomination
Position
Changes
Seats
Address
Work
Days
Cuts
Way
People
Supporters
Voters
Proposal
Plans
Stance
Bill
Calling
0.27
−0.14
0.67
0.11
0.11
0.34
0.47
0.33
0.09
0.11
0.58
0.47
0.11
−0.14
0.43
−0.23
0.42
0.72
0.2
−1.0
−1.0
−0.33
0.62
0.43
0.43
0.56
−0.22
0.27
−0.14
0.43
0.19
0.69
0.5
−0.16
0.3
0.2
−0.14
−1.0
0.27
−0.33
−0.33
−0.14
0.56
0.33

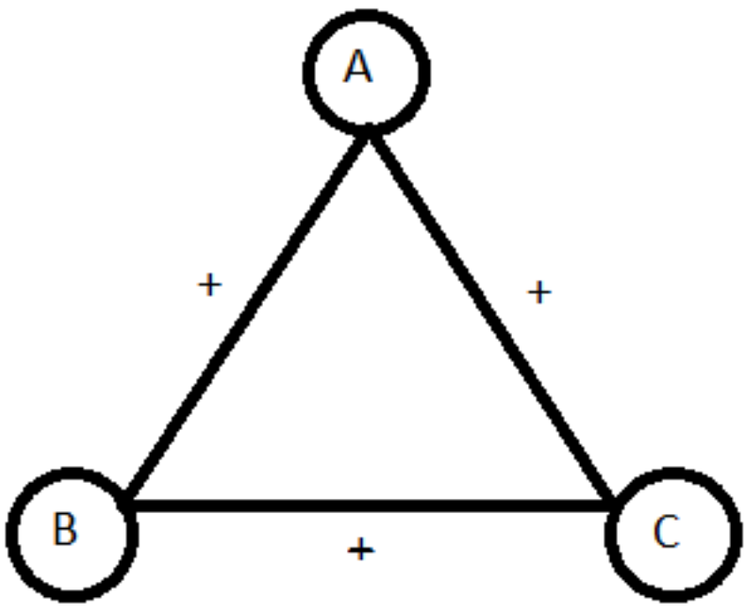


Obama

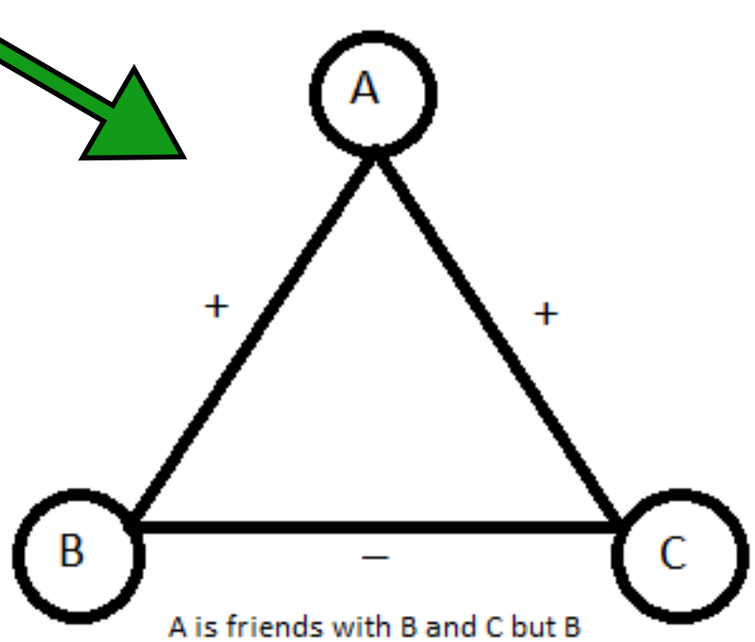
Romney



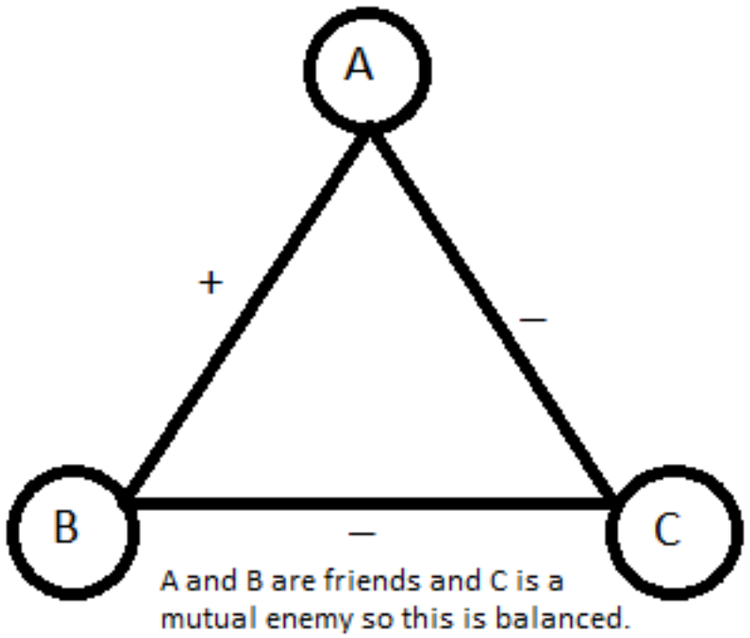




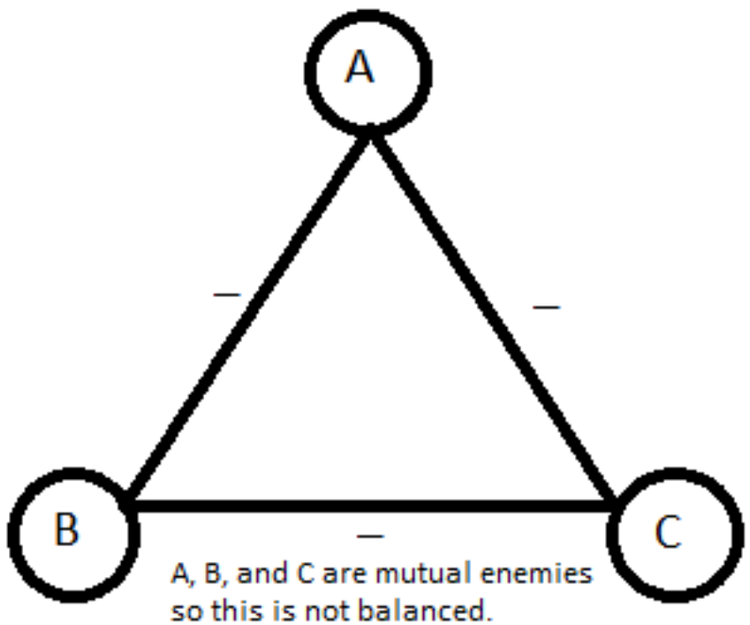
A,B,C all all friends. This is balanced.



A is friends with B and C but B and C are not friends so this is not balanced



A and B are friends and C is a mutual enemy so this is balanced.



A, B, and C are mutual enemies so this is not balanced.

Results indicate interesting implications for structural balance in this particular knowledge network:
There is unbalance that reveals latent points of convergence that are not explicit

Study 1

The Fukushima effect

- The contents of English-language online-news over 5 years have been analyzed to explore the impact of the Fukushima disaster on the media coverage of nuclear power. T

- This big data study, based on millions of news articles, involves the extraction of narrative networks, association networks, and sentiment time series.

- The key finding is that media attitude towards nuclear power has significantly changed in the wake of the Fukushima disaster, in terms of sentiment and in terms of framing, showing a long lasting effect that does not appear to recover before the end of the period covered by this study.

- In particular, we find that the media discourse has shifted from one of public debate about nuclear power as a viable option for energy supply needs to a re-emergence of the public views of nuclear power and the risks associated with it.

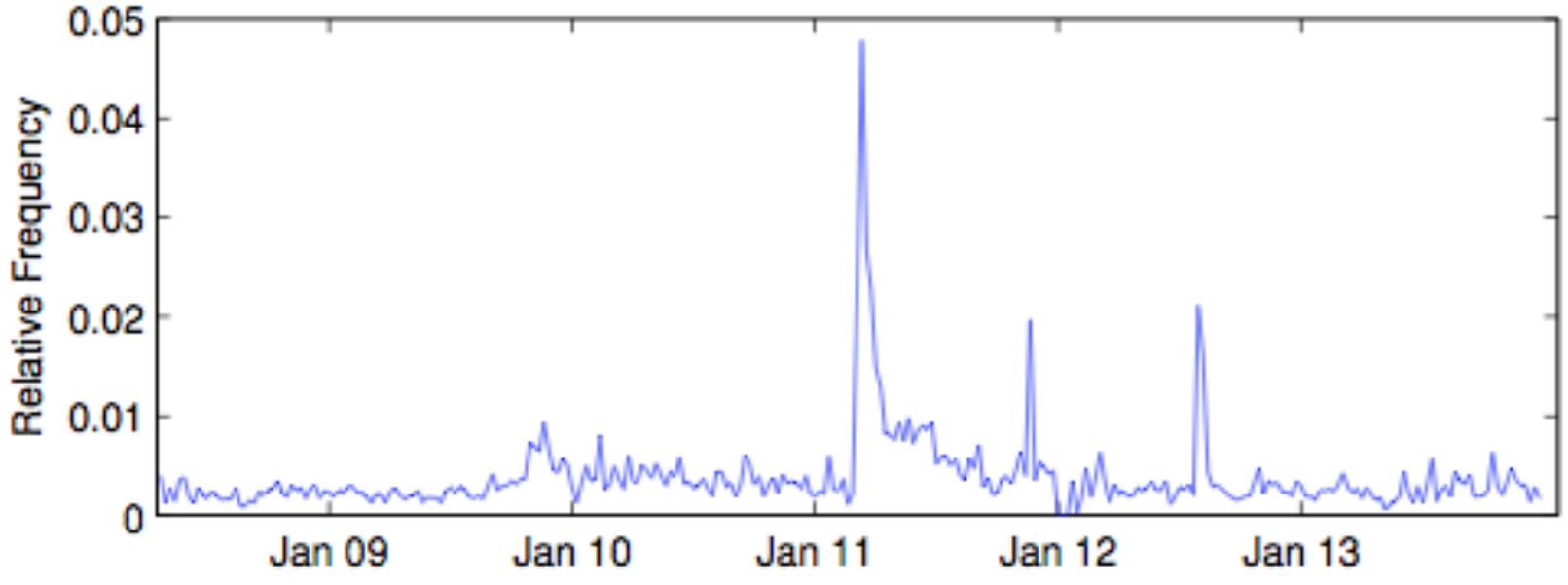


Figure 1. Relative frequency of the number of science articles mentioning 'Nuclear Power' between 1st May 2008 and 31st December 2013.

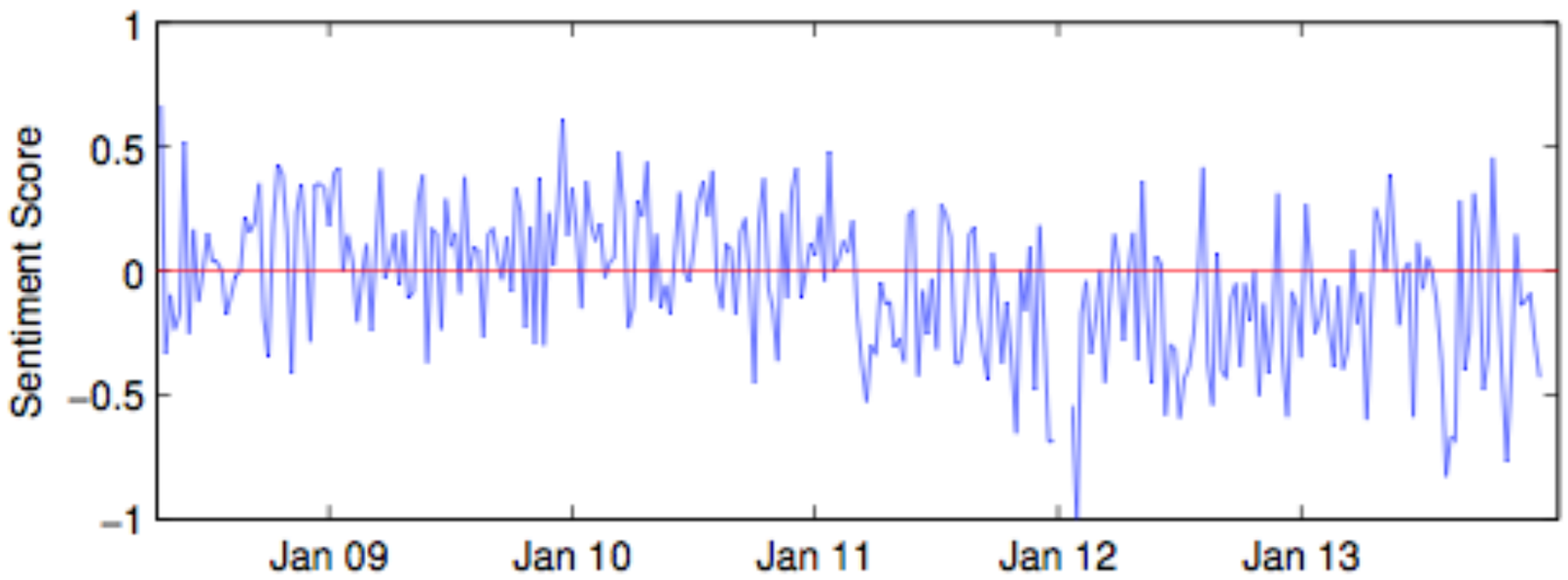
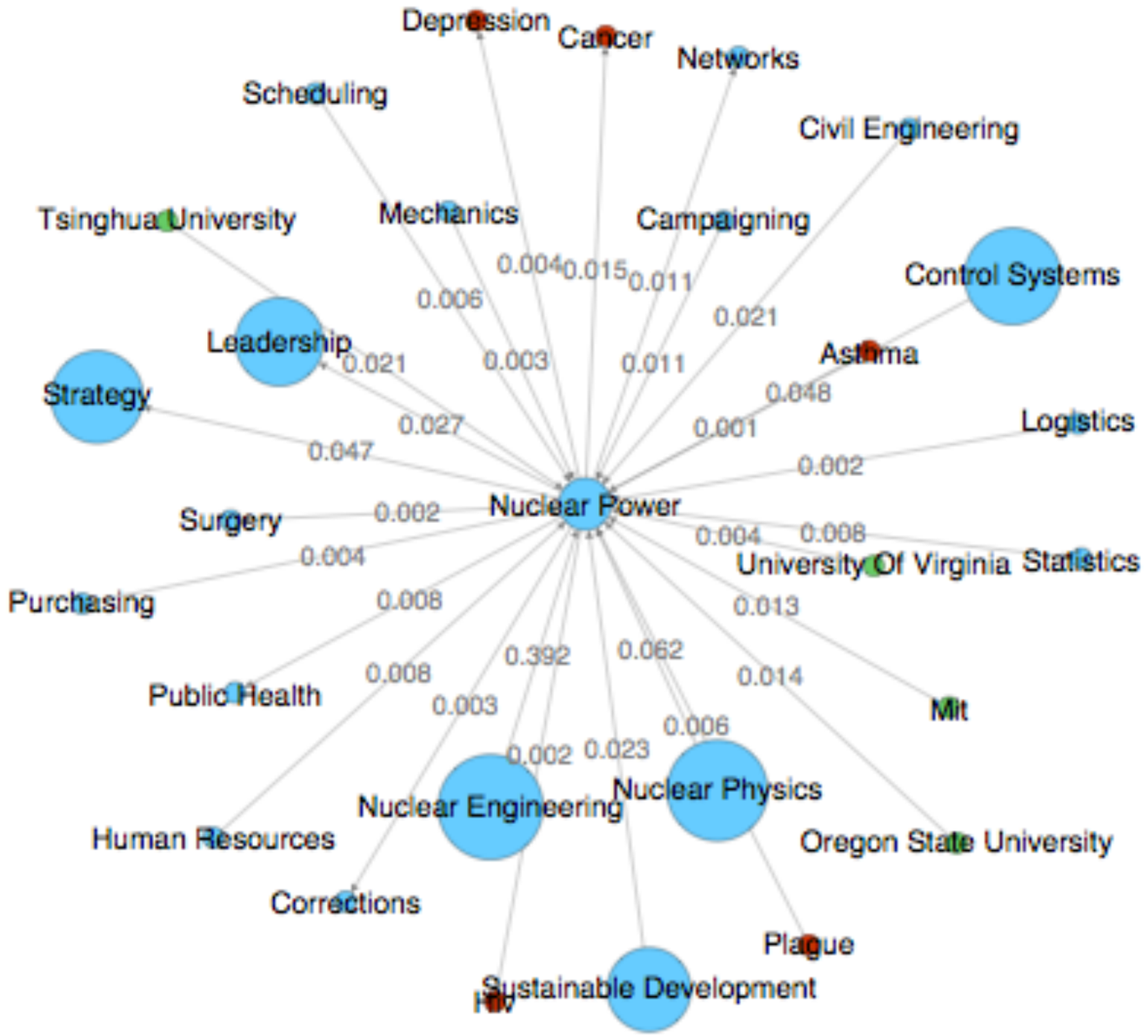
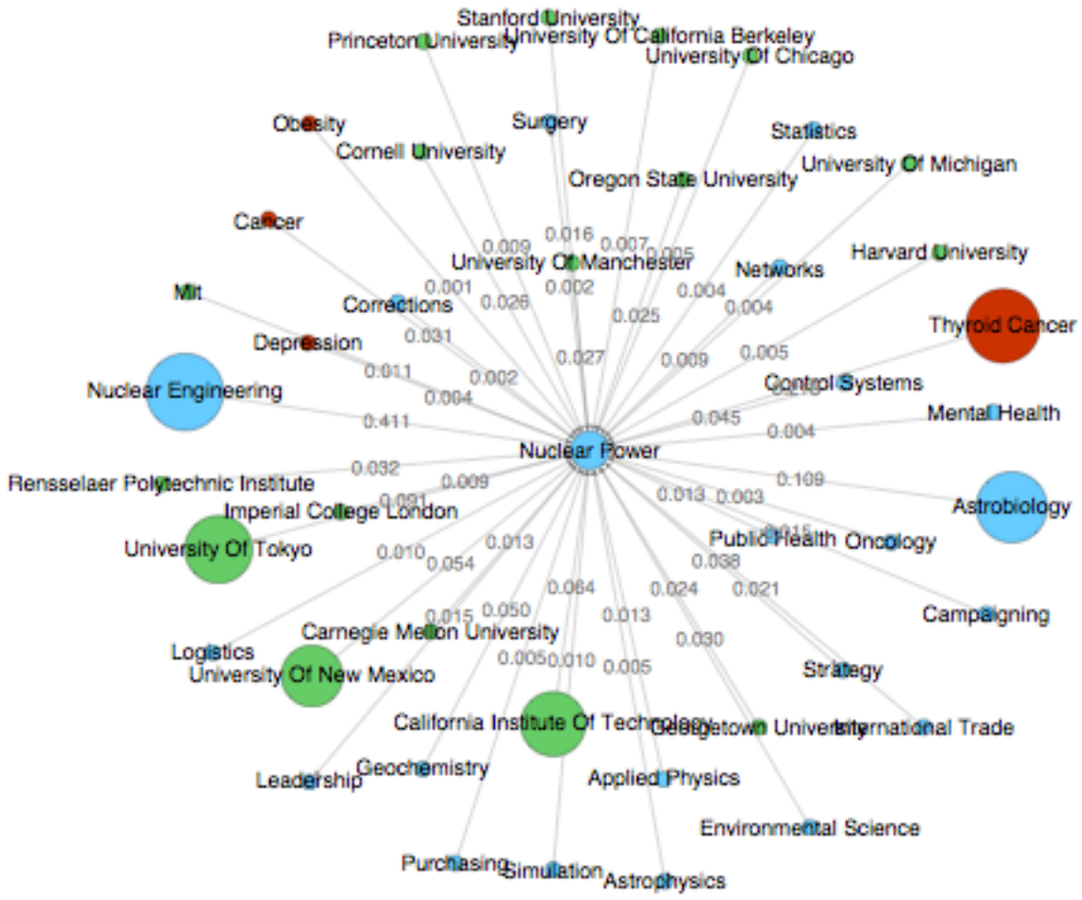


Figure 2. Normalized difference in the number of positive to negative sentences mentioning 'Nuclear Power' between 1st May 2008 and 31st December 2013.


Depression
Cancer
Networks
Scheduling
Civil Engineering
Tsinghua University
Mechanics
Campaigning
Control Systems
Leadership
Asthma
Strategy
Logistics
Surgery
Nuclear Power
University Of Virginia
Statistics
Purchasing
Public Health
Mit
Nuclear Engineering
Nuclear Physics
Human Resources
Oregon State University
Corrections
Plague
Sustainable Development
0.004
0.015
0.011
0.006
0.021
0.003
0.011
0.021
0.048
0.001
0.027
0.047
0.002
0.002
0.004
0.008
0.004
0.008
0.013
0.008
0.392
0.062
0.014
0.008
0.003
0.006
0.002
0.023


Stanford University
Princeton University
University Of California Berkeley
University Of Chicago
Obesity
Surgery
Statistics
Cornell University
University Of Michigan
Oregon State University
Cancer
Harvard University
University Of Manchester
Networks
Mit
Corrections
Thyroid Cancer
Depression
Control Systems
Nuclear Engineering
Mental Health
Nuclear Power
Rensselaer Polytechnic Institute
Astrobiology
Imperial College London
Public Health
Oncology
University Of Tokyo
Campaigning
Carnegie Mellon University
Logistics
University Of New Mexico
Strategy
California Institute Of Technology
Georgetown University
International Trade
Leadership
Geochemistry
Applied Physics
Environmental Science
Purchasing
Simulation
Astrophysics