Obesity and Health over Social Networks: Final Report

Our project aimed at repositioning questions related to environment, behavior and public health to the domain of social network platforms, such as Twitter. It has soon become clear to us that the intended focus on obesity was difficult to be pursued in the short timescale of this funding, as the data retrieved from Twitter contained a significant amount of noise that we were unable to handle. We hence decided to focus on the less endemic life-style related condition of Eating Disorders (namely Anorexia and Bulimia). The reason for doing so is two-fold: first, the nature of the disease makes the technical scope of our analysis a lot more tractable; second, this latter topic shares with our original choice many aspects of the social, rather than just clinical dimension of life-style related illnesses.

Project Report (Outline):

1. We have gained access to Datasift to run a search of terms relevant to the Eating Disorder condition. This allowed us to collect a list of 4,500 persons of interest (POI). Our RA, Brendan Neville, has then developed software that was deployed to directly collect historical data of the last 3,000 items of communication of each of these people, filtered for specific tweets of 'reply-to' (28,993) and 'mentioned' (17,227). By retrieving data about follower-followee and re-tweets relationships, and on locations and demographics, we were able to produce a network of Tweets, on which some preliminary network analysis has been run. Result so far hint at a strong core-periphery structure of the network of interest. There appears to be a core of very active users who receive many messages from and send many messages to other very active users (see attachment).

2. We used NodeXL to perform a search of the tweets in a given week that contained the identified key-words (Anorexic or Bulimic). This returned a network containing 11682 users with 13422 edges. We have represented the information in various graphical forms (grouping by Clauset-Newman-Moore cluster algorithm and laid out using the Harel-Koren Fast Multiscale layout algorithm) to get a preliminary feeling of the clustering of tweets. One of the useful tools in NodeXL allows looking at the content in the tweets, both for the entire graph, and within the groups, by thus providing a conceptual summary of the groups.

3. Through Linguistic Inquiry and World Count (LIWC) we analysed the language used in a sample of tweets collected from our target group (identified by their own profile description). LIWC produces a report on emotional, social and thinking style based on the use of the language. We intend to run this on a larger scale and on key identified groups. Results so far suggest distinct patterns of dialectical interaction of pro-pathology and pro-recovery attitudes.

4. We have become aware that in the social world of Eating Disorders (self-named AnaMia) many user profiles contain detailed quantitative information regarding current and target weight of the users. On the basis of this, we were able to compute users' BMI for a substantial subset of our sample. This finding grants future research.

5. We have produced an extensive bibliography on the topic of Obesity and on the topic of Eating Disorders and we have identified relevant potential liaisons in key academic institutions and international research arena. We have identified interesting research questions that we plan to address in the coming months, compatibly with our other teaching and research commitments. We intend to apply for further funding in order to release our time and make it possible to address these questions in a timely manner.

6. We have secured funding from the ESRC-DTC and from IfLS for a PhD studentship. We are pleased to report that the position has been filled by Tao Wang, who has started working on this topic.

Collaborations with External Stakeholders and Impact:

Collaborations: We have already established connections with a number of national and international experts on the topics of Public Health, as well as experts on the study of Social Network Data. The feedback on the work we want to do on obesity/eating disorders (and in general Life-style related NCDs) is very positive and raises significant interest. We intend to pledge funding to organize a one-day conference here at Southampton as we feel that this initiative could foster our interdisciplinary collaborative approach within a multi institutional framework of interaction between economists, public health experts, computer scientists and general academic users. At present, our liaisons include:

1. A number of economists, e.g. as R. Griffiths (Manchester) and S. Smith (Bristol) work on estimating peer effects with field data. We are interested in applying these econometrics techniques on data sifted from social networks;

2. Prof. Mireia Jofre-Bonet at City University, who is an economist who has done work on Obesity and on Eating Disorders. http://www.city.ac.uk/people/academics/mireia-jofre-bonet/

3. Prof. Donald Kenkel at Cornell University, who is an economist expert in Human Ecology http://www.human.cornell.edu/bio.cfm?netid=dsk10;

4. Prof. Philippe Giabbanelli http://www.dachb.com/, who has now relocated to Illinois.

5. Professor Paul Roderick at the Medical School of the University of Southampton, who is a public health specialist with interest in health behaviors and their social network extensions.

We would like to build new connections with:

6. Academics here at the University of Southampton working on related topics within the Medical School and/or the Health Care profession and the Web Science Institute;

7. The Cathie- Marsh institute in Manchester with expertise on social network data: http://www.socialsciences.manchester.ac.uk/subjects/social-statistics/our-research/centres/cathie-marsh-institute-for-social-research/;

8. Alan Turing Institute(s) for Data Science.

9. M. J. Paul and M. Dredze Human Language Technology Center of Excellence, Center for Language and Speech Processing, at Johns Hopkins University: http://www.hltcoe.jhu.edu/

User Engagement and Brokerage: We feel that our project is at the stage where it could benefit the most from investment in knowledge brokerage activities, such as networking and building relationships with users and/or producers of related research. To this aim, we would like to expand our activities to build (and /or foster) connections with:

1. A charity called Beat, who is engaged in research on Eating Disorders, for example via the administration of surveys http://www.b-eat.co.uk

2. A research center, called ANAMIA involved in the study of Eating Disorders http://www.anamia.fr

Potential Impact: We reckon that further investment in Research Brokerage Activities (for an estimated cost of £25.7k as from an (unsuccessful) bid to the EPSRC Institutional Sponsorship Award) would make it possible for us:

- to up the stakes regarding our project on Obesity and Health over Social Networks, by scoping collaborations and brokerage across multiple institutions and by developing relationships and networks with, among and between producers and users.

- to create a research focus on the topic of the transmission of Life Style related conditions over Social Networks, strengthening the positioning of the University of Southampton.

Obesity over Social Networks: Outline for a Large Scale Grant Proposal (A. Ianni – Lead)

Background:
The latest WHO Global Status Report on non-communicable diseases (NCDs), published in 2014, identifies NCDs as one of the major health and development challenges of the 21st century, in terms of their social, economic and public health devastating impact. Our project aims at contributing to the analysis of how a specific NCDs, namely Obesity, presents a relevant social, rather than just clinical dimension, by highlighting the importance of its online social determinants, among other more traditional forms of creation and maintenance of social ties.

Obesity:
Obesity has manifested itself into an epidemic over the past 30 years and is now a major contributor to the global burden of disease (WHO 2000, Knai et al. 2007), although there is only limited information as to the specific spread patterns (Ejima et al. 2013). Past studies have highlighted the role of peers and social networks, with Blanchflower et al. (2009) suggesting taking seriously the possibility of socially contagious obesity. Similarly, Christakis et al. (2007) quantitatively model the nature and extent of the person-to-person spread of obesity and find that networks significantly influence the biologic and behavioural traits of obesity, which appears to spread through social ties.

Given that health behaviours have been shown to spread through networks, a natural question is whether such patterns can be identified through web-based social media like Facebook or Twitter. This project aims at repositioning questions related to environment, behavior and public health to the domain of social network platforms, such as Twitter. Language traits collected from Twitter have recently been proved to significantly affect economic behaviour (Chen 2013) and are also shown to improve predictive accuracy of demographic statistics (Culotta 2014). The combination of newly developed data analysis techniques for Big Data and twitter-based statistics for sentiment and trend analysis has developed into an established field at the interface of Computer Science, Economics, and the Social Sciences.

Furthermore, twitter-based data analysis has already been shown to provide useful information about health-related questions, e.g. tracing and predicting the spread of infectious diseases like influenza (Lampos and Christianini 2010; Culotta 2010) even leading to the development of Twitter based platforms for Public Health Surveillance (Dredze et al. 2014). Established methods for filtering and analysing health related Twitter data such as the Ailment Topic Aspect Model (Paul and Dredze 2014) exist, but have not yet been used beyond applications for monitoring the (geographical) spread of diseases. In addition, data extracted from Twitter in health related contexts have not been used for individual profiling or for investigations of correlations and spread on social networks.

Here we propose to extend existing methodology to explore the spread of life-style related health conditions over social networks. The research proposes to use data analysis to filter large datasets of Tweets, building individual profiles of Twitter users. The primary focus will be on collecting data related to obesity, but also other factors characterizing individual users will be taken into account. Data gathering methodologies will also be used to infer social networks of Twitter users and this will allow us to verify the earlier controversial survey-based studies of Christakis et al. (2007) using significantly larger datasets. In comparison to Christakis et al. (2007) our approach also enables us to collect and aggregate temporal information which might allow direct conclusions about disease spread, but it also needs to be born in mind that ties in online social networks bear distinct characteristics compared to social ties measured by survey-based techniques, and that the demographics of users of social media does not give a representative cross section of the entire population. Hence, the proposed research involves significant effort in data collection, data filtering, and comparison to datasets from other sources.

More specifically, we aim to address the following questions:

1. Can twitter-based information be used to reconstruct health-related information about individual users? Can we build profiles of individual users in this way and verify survey-based studies on the spread of obesity using large datasets?
2. Can twitter-derived information be embedded into traditional models to improve their predictive power?
3. Can we characterize the way in which obesity spreads over the network, by modelling two counteracting effects: a behavioural effect (anchoring and self-enforcing lifestyle) that leads to contagion and a risk-sharing effect (by which the network provides support)?

4. Can we use Twitter-derived information to design health policies that exploit the explicit network connections, by identifying key players, so that resources invested in health care policy could be efficiently targeted?

We envisage that the project will proceed in several stages:

1. Pilot and data gathering stage: experiments with various ways of data gathering will be explored, aiming to compile a small dataset to explore methodology. We will investigate the use of both direct data gathering using the Twitter API, but also link to commercial solutions (e.g. Datasift) and the expertise of the Web Science Institute. We aim to implement the Ailment Topic Aspect framework of Paul and Dredze (2011) as a first method for data filtering, but also other probabilistic models will be explored.
2. Various ways to construct social networks from our data sources will be explored. One obvious option is via Twitter, which enables us to retrieve data about follower-followee relationships; this will provide an estimate of the structure of the "social network" connecting and influencing Twitter users. Another option is to use information about re-tweeting cascades, which can be gathered via Datasift.
3. The first two steps will result in a dataset, which will allow an initial investigation and refinement of our research questions. Cross validation of our findings with other data sources will be explored (One possible way of achieving such comparison is using geo-information from twitter for comparisons to region specific statistics from the NHS).
4. In a fourth phase the scalability of our methodology to very large datasets will need to be explored. Once this is achieved, network analysis and an analysis of patterns of spread of health characteristics can be investigated.
5. The construction of a micro founded economic model of contagion of life-style patterns of behaviour on a network, to explicitly incorporate a behavioural effect (anchoring and self-enforcing lifestyle) and a risk-sharing effect (by which the network provides support).
6. The study of the evolution of the identified patterns of behaviour over the dimensions of time and space. We conjecture that the modelled local externalities may lead to bandwagons in the dynamics of lifestyle choices, in a pattern that may display clustering of areas where obesity is prevalent.
7. The characterization of the dynamics with which the phenomenon spreads over a network would provide insights as to the optimal allocation of resources, over time and space, that best achieves improvements in public health.
8. By the nature of social networks, the design and the implementation of optimal health policies that rely on social networks may prove to be an efficient alternative to more traditional ways to disseminate information related to life-style diseases.

References:

* Blanchflower D.G. et al. (2009), Imitative obesity and relative utility, Journal of the EEA;
* Chen M.K. (2013), The Effect of Languages on Economic Behavior, American Economic Review;
* Christakis N.A. et al. (2007), The Spread of Obesity in Large Networks over 32 years, NE Journal of Medicine;
* Culotta A. (2014), Estimating County Health Statistics with Twitter, mimeo, Illinois Inst. of Technology;
* Ejima K. et al. (2013), Modelling the obesity epidemic, Theoretical Biology and Medical Modelling;
* Knai C. et al. (2007), Obesity in Eastern Europe, Economics and Human Biology;
* WHO (2000), Obesity: preventing and managing the global epidemic. Report of a WHO Consultation.
* Lampos, V. and Christianini, N. (2010), Tracking the flu pandemic by monitoring the social web. In: AIPR 2nd Workshop on Cognitive Information Processing (CIP 2010)
* Culotta, A. (2010), Detecting influenza epidemics by analysing twitter messages. ArXiv:1997.4748v1.
* Dredze et al. (2014), HealthTweets.org: A Platform for Public Health Surveillance using Twitter, The first AAAI workshop on World Wide Web and Public Health Intelligence (W3-PHI 2014), Quebec.
* Paul and Dredze (2011), You are what you tweet: Analyzing twitter for public health, Fifth International AAAI Conference on Weblogs and Social Media.

Resources:

The proposed project involves aspects requiring expertise in Economics, Health, Computer Science and Network Science.

Expertise in the area of Computer Science is essential to develop tools to harvest information from Twitter and to filter that information to identify health-related aspects and build individual profiles of twitter users. These aspects of the project will be covered by M. Brede and by T. Wang, based in ECS and in Economics respectively.

Expertise in Network Science is essential to the project as it will allow the reconstruction of social networks, gain an understanding of the spread of characteristics over social networks, and use analysis tools to identify key players and carry out advanced network analysis. These aspects of the project will be covered by A. Ianni and M. Brede who both have a track record of research contributions in the field of Economics and Network Theory.

An assessment of the twitter-derived information, cross validation with other datasets, and impact evaluation requires expertise in Economics and Health economics, which will be provided by A. Ianni and E. Mentzakis.

We have already secured funding for a three year PhD on this topic (from the ESRC-DTC and IfLS).

Skills:

Since October 2014 the team has been working on a smaller pilot project analysing Eating Disorders (namely Anorexia and Bulimia), which constitute a less endemic and technically more tractable form of NCD, which shares many aspects of the social, rather than just clinical dimension of life-style related illnesses that are relevant in the study of Obesity. The deliverable of this pilot project constitutes a sound basis upon which the large grant application will build to pursue the technical analysis, and puts us in a well-equipped position to undertake the project.

Alignment with Research Council strategies:

The proposal aligns with several Research Council strategies. For instance, it directly addresses issues related to two challenge areas of the EPSRC Digital Economy theme, i.e. the area of Sustainable society and the area of Communities and culture. Relevance to the first derives from the increasing social and economic cost of treating diseases with direct links to obesity (e.g. diabetes), whereby our results will provide a better understanding of the patterns of spread, and so will enable better treatment and prevention opportunities. Relevance to Communities and culture derives from opportunities to use information to develop strategies that try to target and influence key players.

There are also less direct but nevertheless clear links to other EPSRC themes, like Healthcare Technologies (Can better statistics be used to target treatment?), Global Uncertainties (How can we afford the cost of healthcare in the future?) and to the Physical and Mathematical Sciences themes (with questions related to data processing and data extraction methods).

The proposal is also related to other funding bodies such as the programs on Public Health Research (PHR) and Health Services and Delivery Research (HS&DR) of the National Institute for Health Research (NIHR). Focus is on the evaluation of non-NHS interventions intended to improve the health of the public and reduce inequalities in health (e.g. call on 15/49 Public Mental Health) and respectively on improvement of NHS interventions delivery of services, quality, accessibility and organisation of health services, including costs and outcomes. Finally, a proposal could be aimed at the funding streams of the Leverhulme Trust, which support innovative and original research projects of high quality that potentially lack niche funding streams and bodies.

Costing:

Staff costs (for a minimum of 24 months):

A. Ianni @ 0.2 FTE; E. Mentzakis @ 0.2 FTE; M. Brede @ 0.2 FTE; P. Roderick @ 0.1 FTE; Professor of Health Psychologist @ 0.1 FTE; Research Fellow (Post-doc level - computer science) @ 0.5 FTE; Research Fellow (Post-doc level - network science) @ 1 FTE.

Miscellanea cost:

Infrastructure (PC hardware & software, servers etc) @ £15K; Conference attendance @£20K; Advisory group meeting @ £15K; Organize conference & Dissemination@ £10K.

Short report: Analysis of a communication network of persons of interest with eating disorders as identified by twitter data

Markus Brede[1] , Antonella Ianni[2] and Emmanouil Mentzakis[2]

[1] *Electronics and Computer Science, University of Southampton*
[2] *Department of Economics, University of Southampton*

*** Missing PACS ***

Abstract. - A short presentation of Twitter data analysis of persons of interest with eating disorders and their networks.

Network analysis. – In the following I analyse the network made up of a dataset of 4500 persons of interest (POI). The last 3000 items of communication of each of these people has been collected and filtered for specific tweets of "reply-to" and "mentioned" tweets originating by a POI and being targetted at a POI; we find 46220 such communications of which 28993 are of the "reply-to" type and 17227 'are of the'mentioned type. For the purposes of this report I do not distinguish between both types, but treat them as the same type of communication.

These data can be interepreted as a directed network of targetted communications between POI that might reflect the structure of an underlying social network of acquaintanceships. I see the network as a weighted network, a weight of a link from person A to person B is constructed by assigning it the count of all communications from person A to person B. One may represent this network by a binary adjacency matrix A and an additional matrix W that carries information about the weight of links.

Analysing this network gives us some picture of the structure of communication in the community of POIs. This network is highly heterogeneous, marked by skewed distributions in a number of key quantities. An example is the distribution of link weights in Fig. 1(a) which shows that the probability that two POIs have w communications roughly follows a power law $P(w) \sim w^{-\alpha}$ with $\alpha \approx 2.1$, i.e. few people communicate very intensely, while communications between typical people are of a one-off nature or are only repeated very few times in the timeframe of interest.

This heterogeneity is also reflected in the distributions of in- and out-degrees ($k_i = \sum_j a_{ji}$, $k_o = \sum_j a_{ij}$), cf. Fig. 1(b). In-degrees give the number of unique POIs from which a POI received communication, out degrees give the number of unique POIs a POI has directed communications to. Both distributions show that there are as well nodes that have not sent (or received) communications, as well as well connected senders/receivers. Both distributions have the shape of a power law with exponential cut-off. The node that has sent messages to the largest number of other unique POIs is node (IEDAction, International Eating Disorder Action, ID:2860208794, $k_o = 141$), the node that has received messages from the largest number of unique POIs is node (ID:21700435, aedweb The Academy for Eating Disorders, $k_i = 446$).

The picture can be refined by also taking account of the numbers of communications sent which is reflected in the distributions of in/out-strengths ($s_i = \sum_j a_{ji} w_{ji}$, $s_o = \sum_j a_{ij} w_{ij}$), cf. Fig. 1(c). One again notes distinct heterogeneity in the distributions of in- and out strengths, reflected in the linear behaviour in log-log plots. There are no apparent exponential cut-offs, but one might speculate about a cross-over between two power law regimeswith exponents around -1.2 and -1.9 which might indicate a natural way to partition the community of POIs into very active members (received and sent more than around 50 communictions) and less active members. We can again identify the nodes with most sent and received communications which are node (ID: 873722743, skinnyanorexic, $s_o = 1411$) and node (ID:2895013596, work_for_thin, $s_i = 3757$), respectively. It is interesting to note that the unweighted metrics pick up organisations as prominent nodes, whereas the weighted measures are maximised by genuine users (I checked this in the list of users).

A question of interest are correlations between the binary adjacency matrix and the matrix of weights of communiction. Exploring such correlations allows us the answer questions as: do POIs who communicate with many

persons of interest also communicate with them very often? To address this we plot the relationship between average in-(out) strength and in-(out) degree, cf. Fig. 2. In fact we find a very strong positive correlation, well described by a power law relationship with exponent larger than one, indicating that POIs with many communication partners communicate significantly more than average with an average partner.

One might also wonder how in- and out-degrees (or strengths) are correlated. Do POIs who send (many) messages to many others also receive (many) messages from others? We analyse this relationship by plotting the average in/out degree vs out/in degree in Fig. 3. The answer is again positive: up to a saturation point of around 100 communication partners POIs with more partners who send messages to them also tend to have more partners who receive messages from them. This positive correlation is again well characterised by a power law (with exponent again around 1.2). As expected from the positive correlation between degree and strength, this picture is confirmed when analysing the dependence of average out/in strength on node degree (also Fig. 3).

Next, we explore degree mixing patterns, are nodes with high degree linked to other nodes of high degree? Or similarly with strength? On can explore this question by defining average neighbour degrees via $\langle k_i^{in,nb} \rangle = 1/k_i^{in} \sum_j a_{ji} k_j^{in}$ and then plot the average dependence of the average degree of a neighbour on the degree of a node. This plot is provided in Fig. 4 where we see that average in/and out degrees of neighbours of a node initially grow with the in-degree but then decay. Hence, a node with low or large in-degree tends to be linked to average nodes (in case of a node with large in-degree this makes sense, in so far as that such a node is connected to a large fraction of the population that one expects that it sees the average individual), but nodes with intermediate in-degree tend to form connections to nodes of large in and out degrees.

This picture is modified quite crucially when consider strength-strength correlations and accounting for the weighted nature of links in these correlations. Average neighbour strength would then be defined via $\langle s_i^{in,nb} \rangle = 1/s_i^{in} \sum_j a_{ji} s_j^{in}$, which is explored in the bottom panel of Fig. 4. Here we see strong signs of strong assortative mixing: nodes who send and receive many messages tend to be connected with similar nodes.

These findings so far hint at a strong core-periphery structure of the network of interest. There appears to be a core (maybe defined as in Fig. 1(c) of very active users who receive many messages from and send many messages to other ver active users. A periphery is made up of the opposite type of user

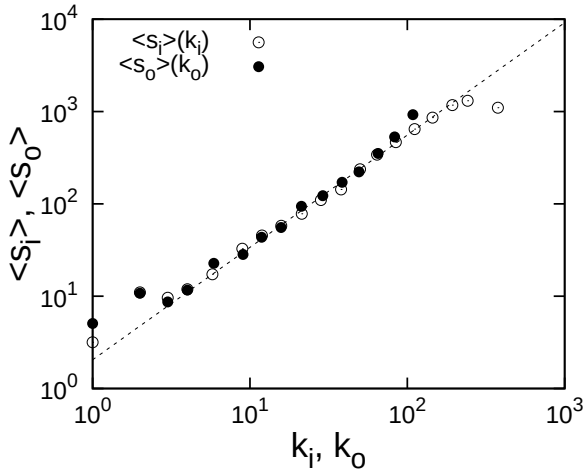

Fig. 2: Dependence of node strength on node degree. One finds a strong positive correlation characterised by the power law relationship $s \sim k^\gamma$ with $\gamma = 1.21 \pm 0.05$.

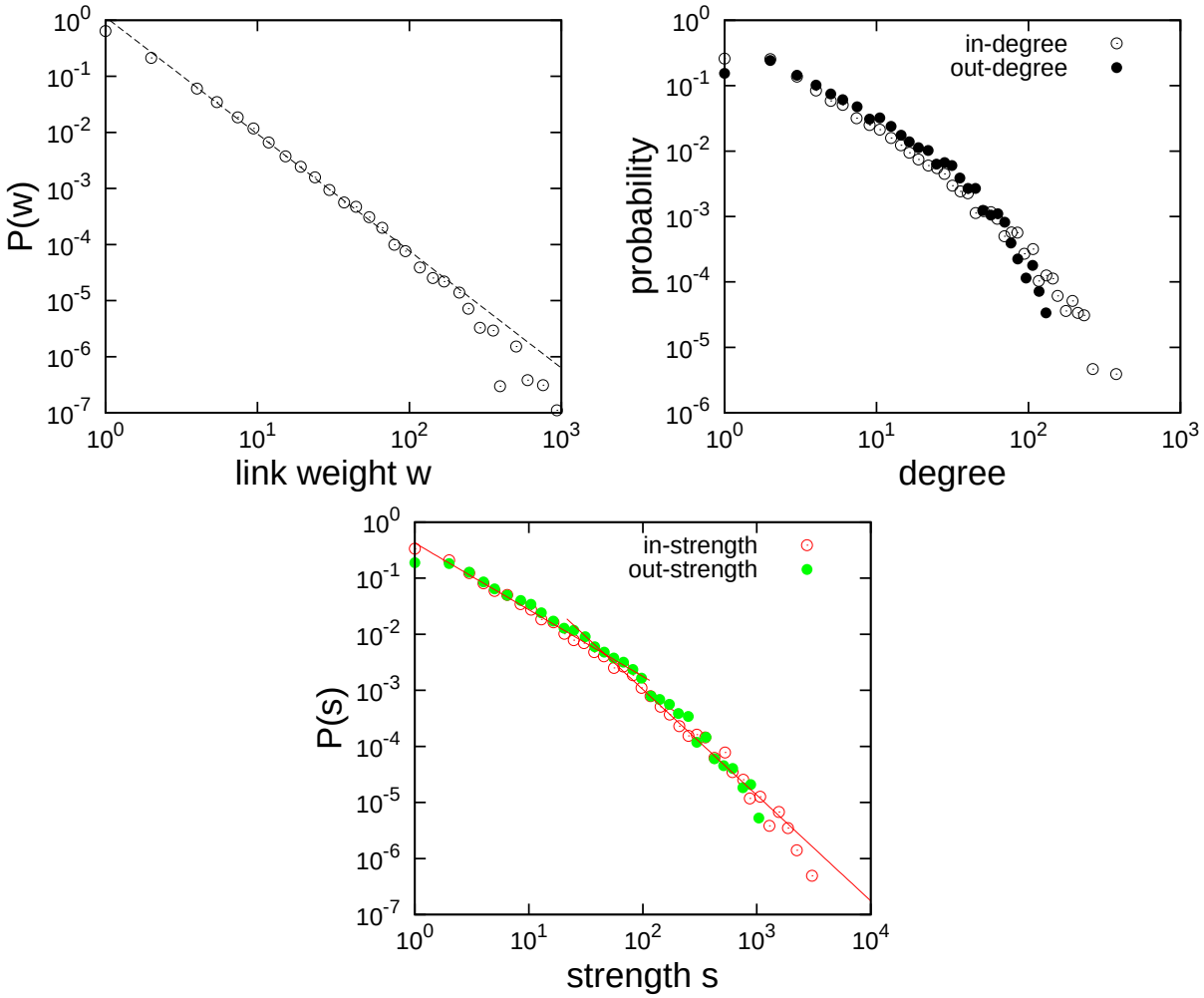

Fig. 1: (a) Distribution of link weights. The line indicates the best fit of a power law with exponent -2.1 ± 0.1. (b) The distribution of in and out-degrees (actually of degree plus one to include nodes with degree zero in the log-log plot). (c) Distribution of node strength (plus one). The lines indicate two possible regimes marked by power laws with exponents -1.2 ± 0.1 and -1.9 ± 0.1 and a possible crossover point around strength 50 (Note that all data have been binned logarithmically)

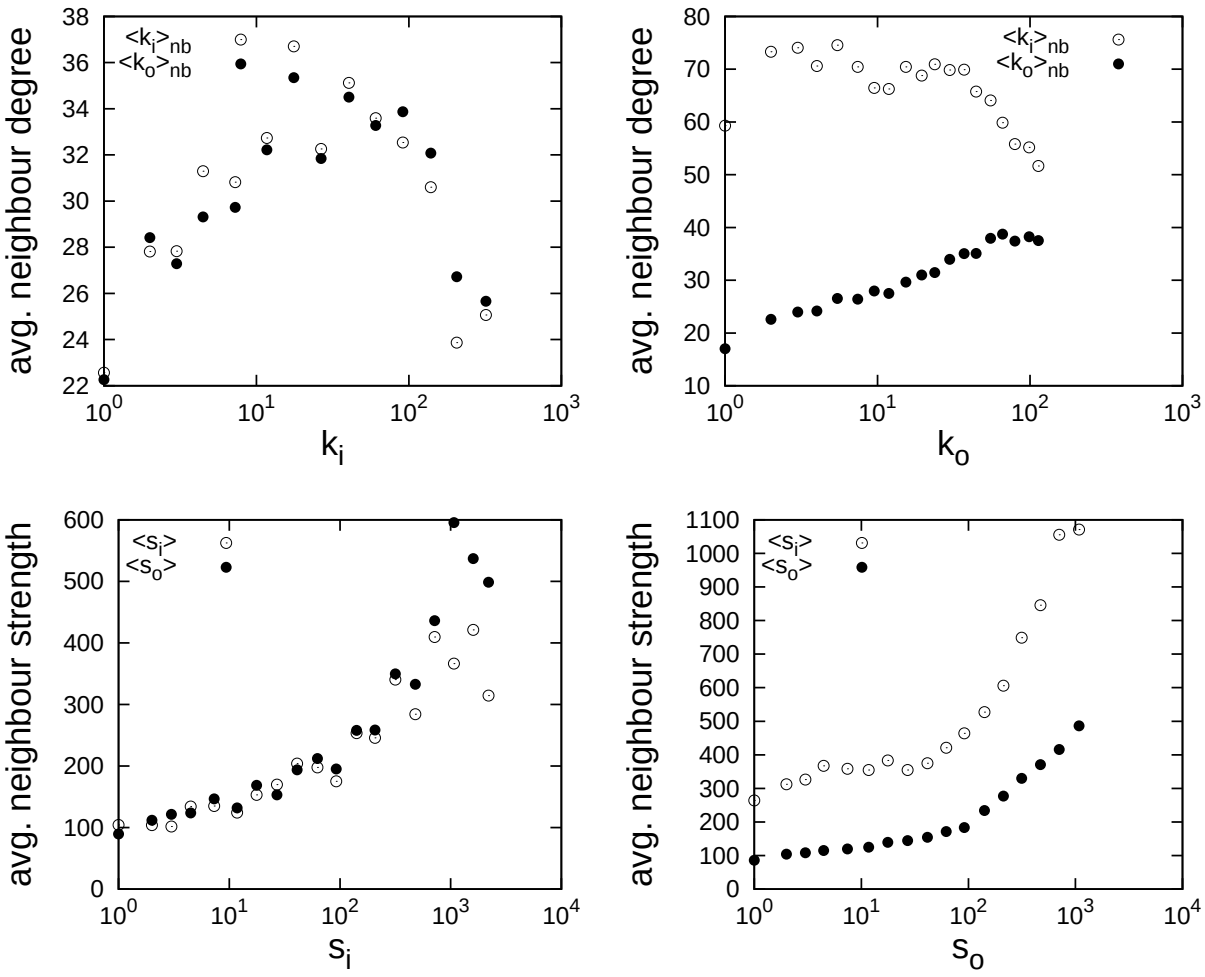

Fig. 4: (top) Dependence of the average neighbour degree on the in-degree (a) or out-degree (b). (bottom) Dependence of the average neighbour strength on the in-strengths (c) or out-strengths (d).(data logarithmically binned)

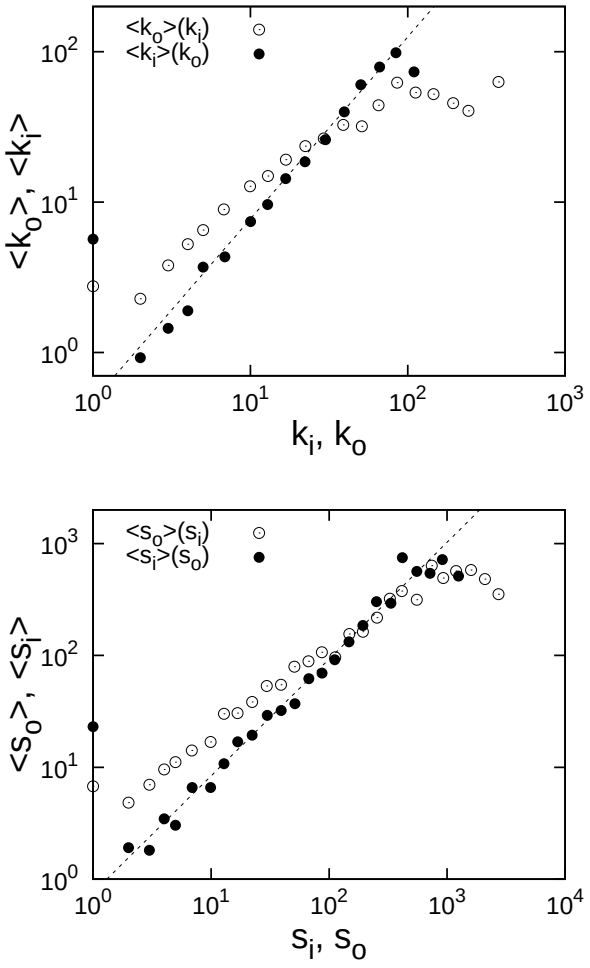

Fig. 3: (a) Dependence of average out/in-degree on in-out-degree. (Excepting degree zero nodes and the nodes with largest out-degree) average in-degrees fit well with a power law with $\langle k_i \rangle \sim k_o^{\beta}$ with $\beta \approx 1.2 \pm 0.1$. Average out-degrees also initially grow strongly with in-degree, but saturate with large in-degrees. (b) Similar analysis to (a), but using node strengths instead of degrees. We find $\langle s_i \rangle \sim s_o^{\beta}$ with $\beta \approx 1.04 \pm 0.1$.