Uncovering the Dynamics of Ego Networks of Scientific Gems

Pietro Della Briotta Parolo[a], Santo Fortunato[a]

[a] *Complex Systems Unit, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland*

Abstract

A promising way to keep track of the impact of a scientific work is to investigate the structure of its ego-network, i. e., of the network consisting of all papers citing that work and their mutual citations. Here we study the dynamics of ego-networks of highly cited articles, and find that it has some peculiar general features. Partial ego-networks, whose papers are published within a sliding time window, are usually very compact in the first years after the publication of the ego-paper, while they eventually fragment in many disconnected components. Their average size peaks after 6-7 years since the publication of the ego-paper and then it steadily decays. These results indicate that in most cases a highly cited paper starts losing visibility within a few years from publication, and its impact is reflected in the success of followup works. The fragmentation of the ego-networks may be due to an increased specialisation of the field of the ego-paper, or a growing popularity of the paper in different fields.

Keywords: Ego Networks, Citation count, Scientometrics

1. Introduction

In social network analysis an ego-network (EN), or ego-centered network, is the graph formed by the neighbours of a specific individual (the ego) and by their mutual relationships [1, 2, 3, 4, 5, 6, 7, 8].

The people in the EN are the ones having the greatest impact on the life of the ego, influencing his attitudes, norms, values, goals and perceptions of the world. Moreover, they are the ones to whom the ego must turn to seek information, help and support. ENs are thus a useful tool to look at social networks from a local perspective.

The concept can be exported to other contexts. For instance, the EN of a scientific paper is the set of all papers citing it, along with their mutual citations. Just like social ENs allow us to uncover the social world of single persons, we can use citation ENs to investigate the impact dynamics of a paper, which is the goal of this work.

We consider ENs of highly cited papers, and study how their properties change in time. To study the dynamics we focus on subsets of each EN, consisting of all papers published in sliding time windows, and their mutual citations. We also investigate the evolution of the full network.

2. Material and methods

2.1. Data description

Our data set consists of all publications (articles and reviews) written in English till the end of 2013 included in the database of Thomson Reuters (TR) Web of Science. We selected recent, highly cited papers, i. e., published after 1990. In the main text we will focus on papers with at least 1000 citations in the first ten years, which add up to more than 2000 papers. In the Appendix we shall vary the threshold to check the robustness of the observed patterns. For each paper, we built the EN of the original work, by selecting the citing papers and returning the citations between them. The networks are created in windows of size w, which indicate how many consecutive years of scientific publications are considered in the analysis.

Fig.1 shows the EN for the famous paper by Barabási and Albert on preferential attachment in complex networks [9]. All the nodes represent papers citing it, along with the connections (citations) between them. A

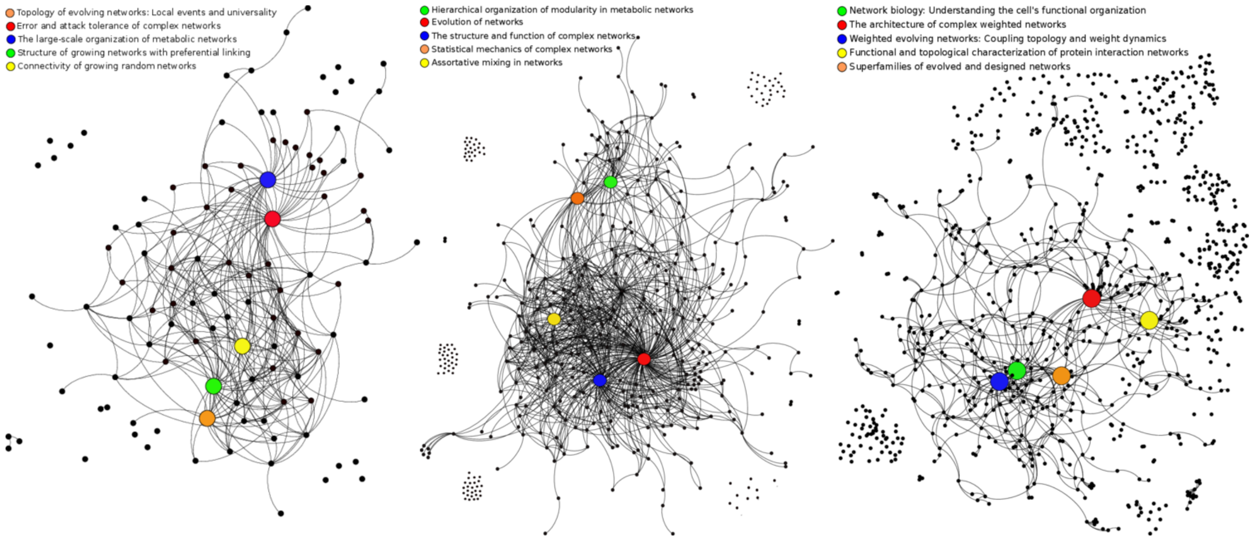
🟧 Topology of evolving networks: Local events and universality
🟥 Error and attack tolerance of complex networks
🟦 The large-scale organization of metabolic networks
🟩 Structure of growing networks with preferential linking
🟨 Connectivity of growing random networks
🟩 Hierarchical organization of modularity in metabolic networks
🟥 Evolution of networks
🟦 The structure and function of complex networks
🟧 Statistical mechanics of complex networks
🟨 Assortative mixing in networks
🟩 Network biology: Understanding the cell's functional organization
🟥 The architecture of complex weighted networks
🟦 Weighted evolving networks: Coupling topology and weight dynamics
🟨 Functional and topological characterization of protein interaction networks
🟧 Superfamilies of evolved and designed networks

Figure 1: Ego-network for the paper *Emergence of Scaling in Random Networks* (Science 286, 509-512, 1999), by A.-L. Barabási and R. Albert. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping and cover the intervals 1-2, 3-4 and 5-6 (years after publication). The EN is initially well connected, its link density is highest at t=3, but it quickly becomes sparse, with a growing number of isolated nodes. Some well known papers are highlighted with colors, their titles are reported at the top.

quick analysis shows some general features of EN evolution: in the beginning they are very dense (left figure), but with the emergence of isolated nodes after a few years (central figure); as time goes by connectivity decreases dramatically and most of the network is made up of isolated nodes and a relatively small connected core.

3. Results and discussions

Properties of the Network

Fig. 2 shows the distribution of EN size for two different window sizes, w=2 (top) and w=3 (bottom) and at different stages in time. The early distribution for the first complete window (the year of publication is not part of it) shows a peak at around 200 ($w = 2$) or 300 ($w = 3$), indicated by the dark red line. Then in the following years the peak moves right (coherently with an increase in citation volume), followed by a retreat until in about 10/12 years the distribution looks similar to the earliest ones. After that, the peak keeps shifting to the left, indicating a decrease in citation volume and a decay of the attention towards the ego-paper [10, 11, 12, 13]. The width of the distributions instead increases with time.

The time evolution of the average EN size can be seen in Fig. 3, for the partial ENs [left panels: w=2 (top), w=3 (bottom)] and for the full one (right panel). For the partial ENs there is an early increase in the mean and median values, followed by a rapid decrease after a broad peak at around $t = 7$. The median falls faster than the mean, suggesting that many networks become small, while a few remain still large. On the other hand the size of the full EN can only increase. The figure shows that the growth is roughly linear in time.

Next, we focus on the structure of the ENs. The first question is whether the network is connected or fragmented into smaller pieces. Figs. 4 and 5 attempt at providing an answer by showing the distributions at different stages of the relative size of the largest connected component (LCC), along with their time evolution. For the partial ENs, the relative size of LCC has initially a rather flat distribution (Fig. 4), so there is a broad range of values that the initial peak can reach. Then the distributions become more and more peaked and shifted to the left, indicating a shrinking of the LCC. The time evolution of the averages can be seen in Fig. 5. In this case there is a more stable pattern, with values starting off around 0.5 and

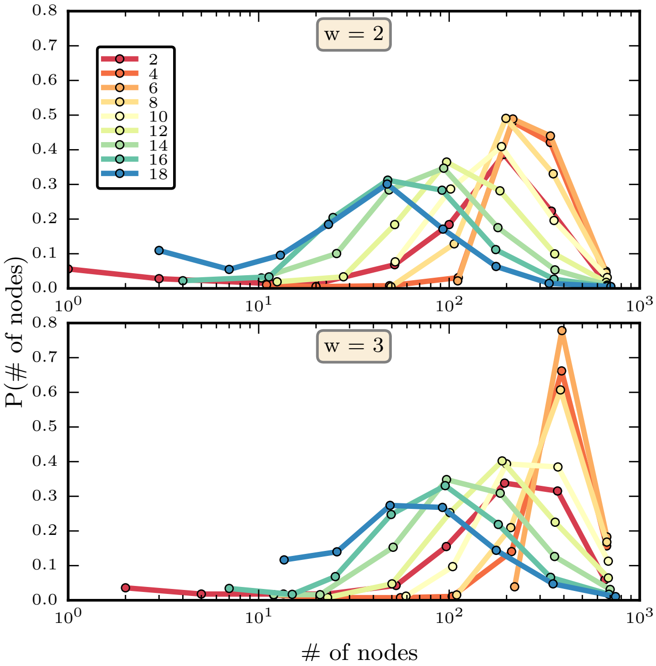

Figure 2: Distribution of EN size, with windows of 2 years (left) and 3 years (right). The different lines indicate different intervals t from the publication of the ego.

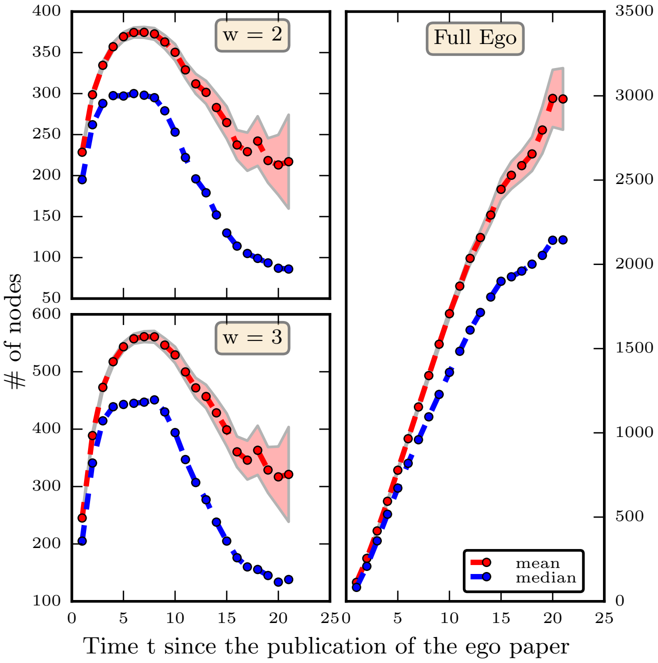

Figure 3: Evolution of mean and median absolute size of the ENs. We also show the standard deviation of the sampled mean. The panels correspond to $w = 2$ (top left), $w = 3$ (bottom left) and to the full EN (right).

then rapidly falling to a plateau which depends on w. For the full EN, as expected, the LCC grows steadily in its initial years, stabilizing at high values after 5 to 10 years from publication. If put together with Fig. 3 this shows that, after the initial transient, papers cite papers in the LCC at an approximately constant rate and the overall connectivity remains strong. Hence we see for the initial years ($2 \leq t \leq 10$) two contrasting phenomena for the partial ENs: an expansion and subsequent contraction of the size of the network and a constant contraction of the relative size of the LCC. What about the *absolute* size of the LCC? Fig. 6 shows the time evolution of the absolute size of the LCC. We can see a very interesting pattern here. After a few initial years of increase (more required for larger w) the mean absolute size of the LCC falls exponentially until $t \approx 15$, where the data starts becoming noisy due to the lower number of papers citing the ego. The result indicates that, as time goes by, recent papers are less likely to cite each other. Further analysis suggests (Figs. A1 and A2) that the collapse of the largest connected component in the window scenario is not associated to fragmentation, but rather to the disintegration of the network. Even though the relative size of the second largest component grows, its absolute size shows a small increase within the first 10 years, followed by a decreasing trend. The full EN consists almost entirely of the LCC, the rest of the nodes forming very small components. Also, Fig A5 shows the impact of the number of citations of the ego on the exponential decay.

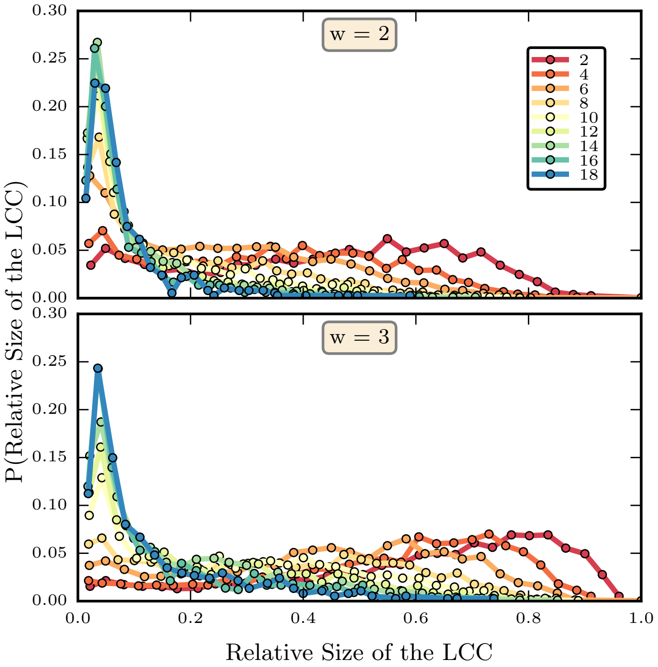

Figure 4: Distribution of the relative fraction of the largest connected component of the EN for all papers with windows of 2 years (top) and 3 years (bottom). The different lines indicate different distances t from publication.

Furthermore, one can look at properties of individual nodes (see also Appendix Figs. A3 and A4) that further characterize the disintegration of the network and of the LCC. Fig. 7 shows the time evolution of the fraction of nodes with incoming degree $k_{in} > 0$, i.e. receiving at least one citation. Consistently with what we have seen before, less and less papers manage to gather any citation within the chosen time window. New papers tend more likely to attach, if at all, to older papers. This can be seen in the full EN, where the fraction of papers receiving citations saturates initially, but keeps slightly increasing. This means that papers keep receiving citations from the other papers of the EN, but not from those within the same time window. Similarly, Fig. 8 shows the time evolution of the fraction of nodes with outgoing degree $k_{out} = 0$, i. e. citing no paper of the EN. This conferms what could be suggested from the previous figure, as on one hand we see that a high fraction of nodes does not contribute citations within their time window, while the total EN keeps receiving citations from the new nodes.

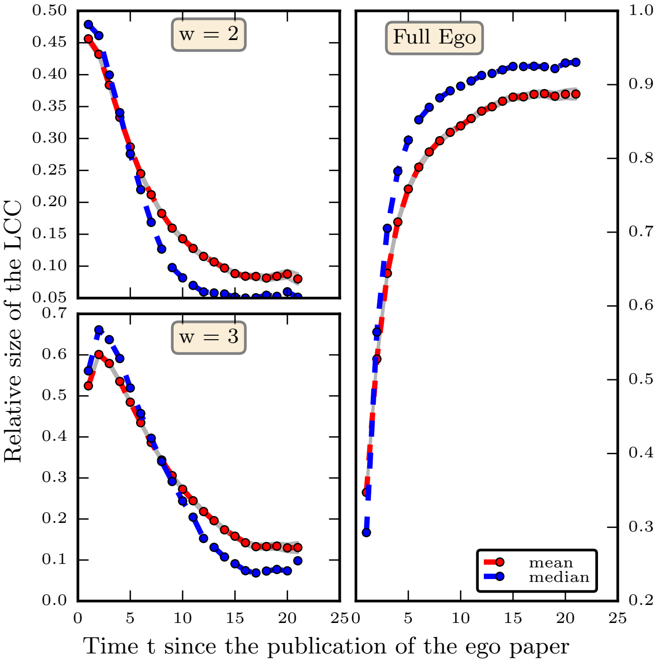

Figure 5: Time evolution of mean and median relative size of the LCC along with the standard deviation of the sampled mean for $w = 2$ (top left) and $w = 3$ (bottom left) and of the full EN (right).

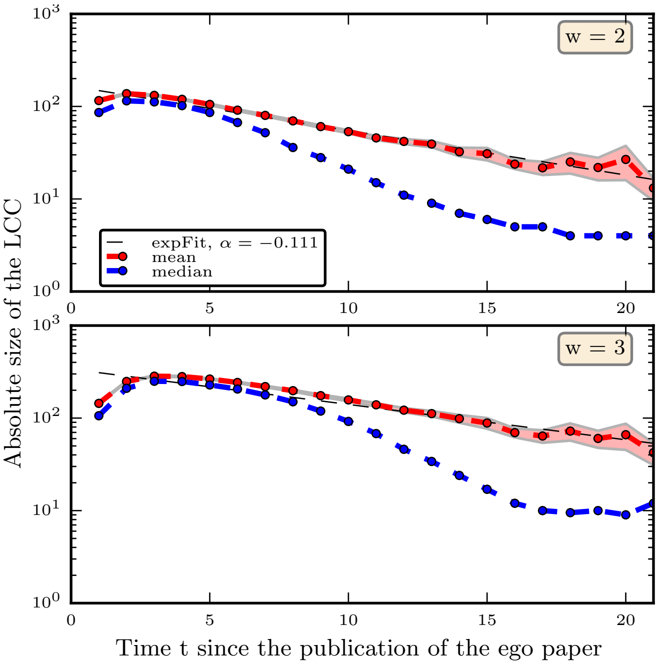

Figure 6: Time evolution of mean and median absolute size of the LCC for $w = 2$ (top) and $w = 3$ (bottom) along with the best fit to the exponential $t = \beta \exp(-\alpha t)$.

Finally, one can look at the relationship between the EN and the overall citation network in order to analyze the impact of the ego paper within its "community". When a new layer of nodes is introduced, each node provides d_i new links. These are just a fraction of the total number of references r_i of the paper. Hence, we can calculate for each paper the value of the fraction $f_i = \frac{d_i}{r_i}$, which quantifies what portion of the reference list goes to members of the EN. Fig. 9 shows the time evolution of the average of this property. As we can see the number initially increases, reaching the peak around 6/7 years after publication, then it decreases. Fig. A6 in the Appendix shows how the number of citations of the ego affects the shape of the curve.

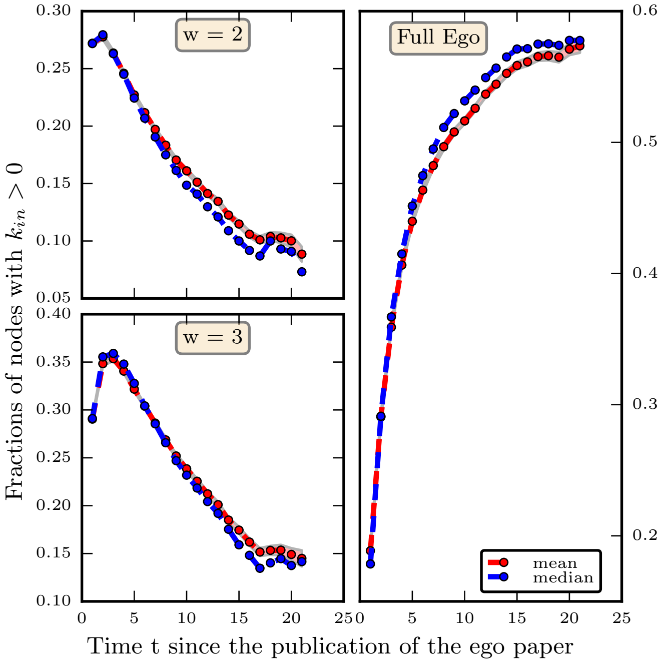

Figure 7: Time evolution of mean and median fraction of nodes with at least one incoming connection from the other nodes of the EN ($k_{in} > 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

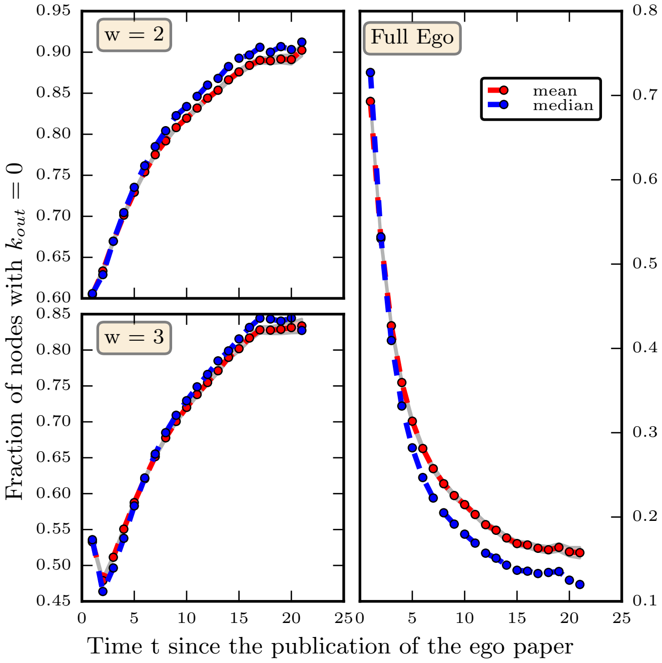

Figure 8: Time evolution of mean and median fraction of nodes without outgoing connections to the other nodes of the EN ($k_{out} = 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

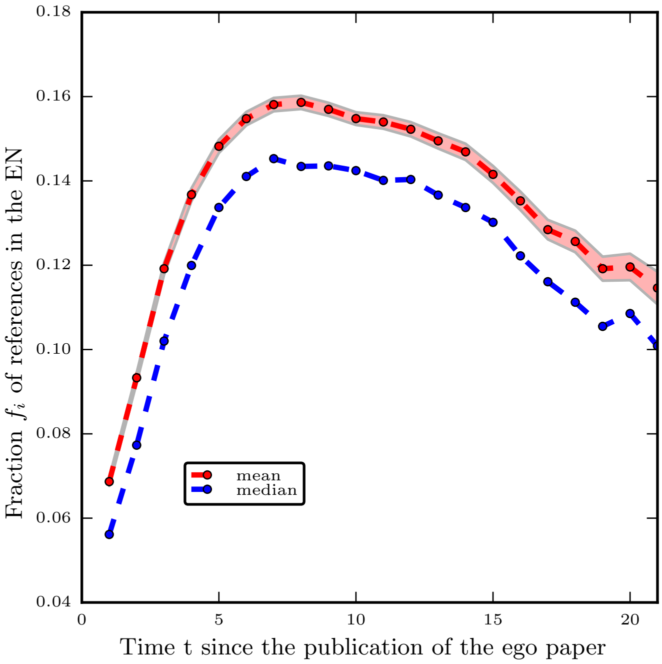

Figure 9: Time evolution of the mean and median of the fraction f_i of references of papers of the full EN belonging to the EN as a function of the number of years since publication.

4. Conclusions

Ego-networks of papers could help us investigate the impact that scientific works have in the literature. We focused on partial ENs, comprising papers published in sliding time windows. We find a consistent scenario, in which the networks fragment into many small components few years after the publication of the ego-paper. The progressive decrease of citations between later members of the EN may signal a specialization of the topic and (or) an increasing popularity of the ego in different disciplines, where citations are infrequent between works on different subjects.

A natural next step of this investigation is proposing a model that describes and possibly predicts the evolution of ENs. Candidate models could build upon popular models of growth of citation networks [14, 15].

Acknowledgments

We used data from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 2013. This research is supported by the European Community's H2020 Program under the scheme ÍNFRAIA-1-2014-2015: Research Infrastructures', grant agreement 654024 *SoBigData: Social Mining & Big Data Ecosystem*, http://www.sobigdata.eu.

Author Contributions

Both authors designed the research and participated in the writing of the manuscript.

References

[1] E. Bott, Family and social network: Roles, norms and external relationships in ordinary urban families, Tavistock Publications, 1957.

[2] L. C. Freeman, Centered graphs and the structure of ego networks, Mathematical Social Sciences 3 (3) (1982) 291–304.

[3] P. D. Killworth, E. C. Johnsen, H. R. Bernard, G. A. Shelley, C. McCarty, Estimating the size of personal networks, Social Networks 12 (4) (1990) 289–312.

[4] S. Wasserman, K. Faust, Social network analysis: Methods and applications, Vol. 8, Cambridge university press, 1994.

[5] M. E. Newman, Ego-centered networks and the ripple effect, Social Networks 25 (1) (2003) 83–95.

[6] J. Scott, Social network analysis, Sage, 2012.

[7] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Analysis of ego network structure in online social networks, in: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom), IEEE, 2012, pp. 31–40.

[8] J. J. McAuley, J. Leskovec, Learning to discover social circles in ego networks., in: NIPS, Vol. 2012, 2012, pp. 548–56.

[9] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[10] A. Avramescu, Actuality and obsolescence of scientific literature, Journal of the American Society for Information Science 30 (5) (1979) 296–303.

[11] T. Pollman, Forgetting and the ageing of scientific publications, Scientometrics 47 (1) (2000) 43–54.

[12] H. Bouabid, V. Larivière, The lengthening of papers life expectancy: a diachronous analysis, Scientometrics 97 (3) (2013) 695–717.

[13] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, S. Fortunato, Attention decay in science, Journal of Informetrics 9 (4) (2015) 734–745.

[14] Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, PloS one 6 (9) (2011) e24926.

[15] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, Science 342 (6154) (2013) 127–132.

Appendix

Further network properties.

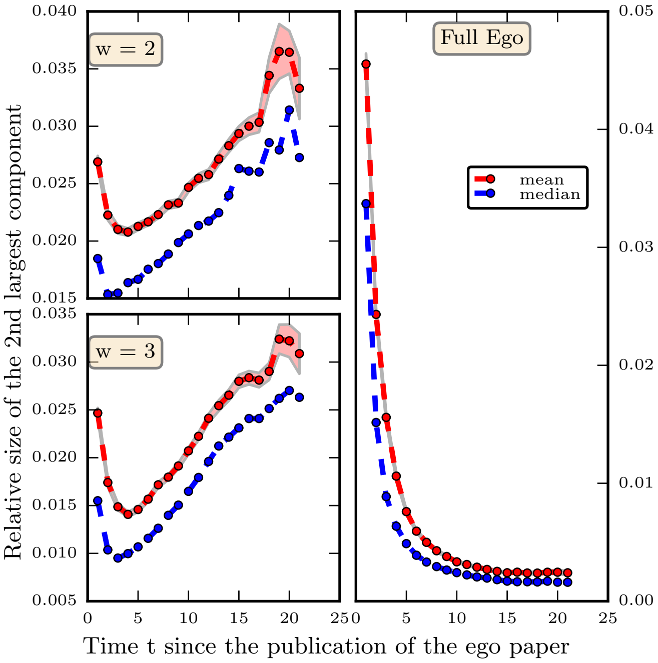

Figure A1: Time Evolution of mean and median relative size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

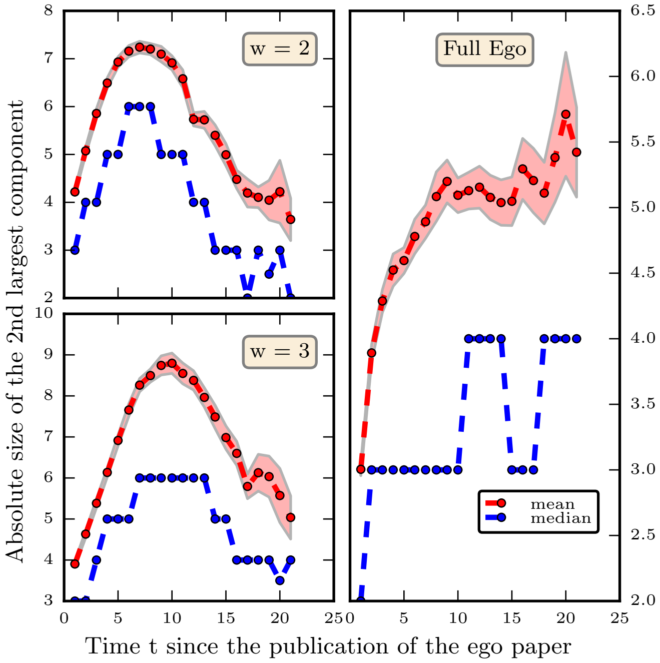

Figure A2: Time Evolution of mean and median absolute size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

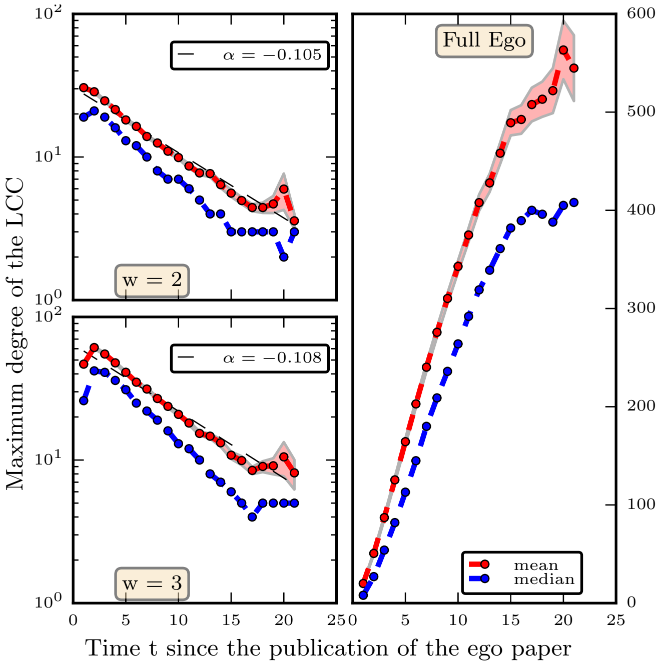

Figure A3: Time Evolution of mean and median relative size of the maximum degree of the lcc for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

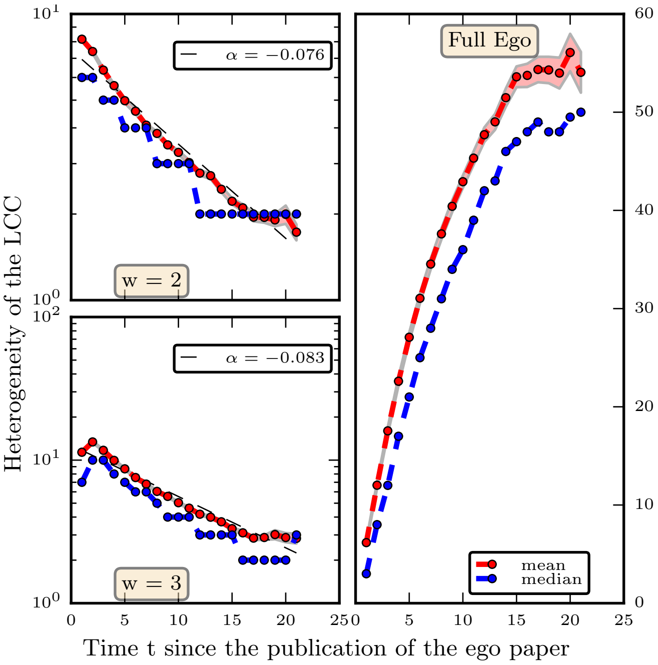

Figure A4: Time Evolution of mean and median relative size of the degree heterogeneity of the lcc, definied as $\frac{\sum d^2}{\sum d}$ for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

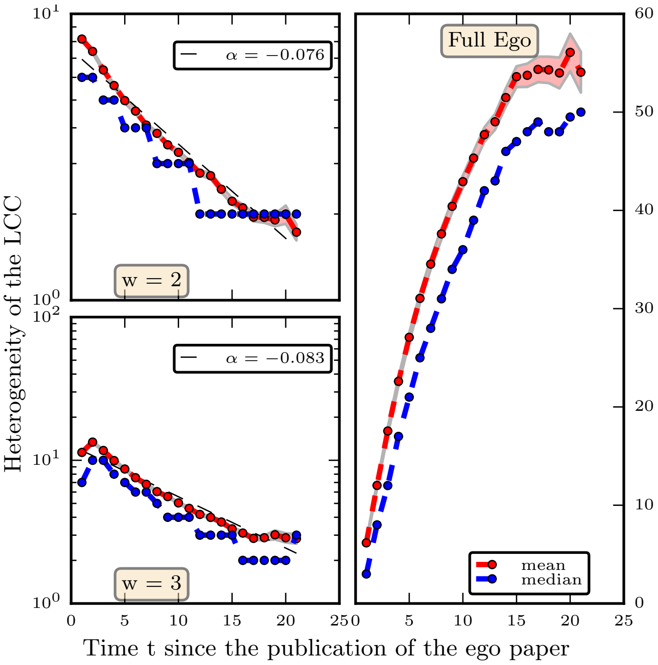

Figure A5: Time Evolution of mean and median relative size of the absolute size of the lcc for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right). he higher the citation volume, the faster the decay. This seems to be caused by the fact that the time required for the network to collapse is identical and thus curves starting from higher values (linked to higher citation volumes) inevitably need to fall faster.

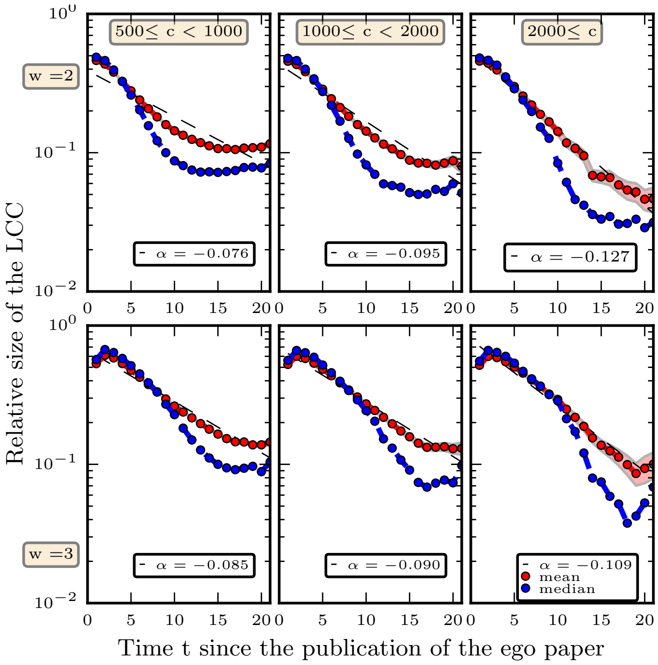

Figure A6: Time Evolution of mean and median relative size of the absolute size of the fraction of references that stay within the EN for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right).