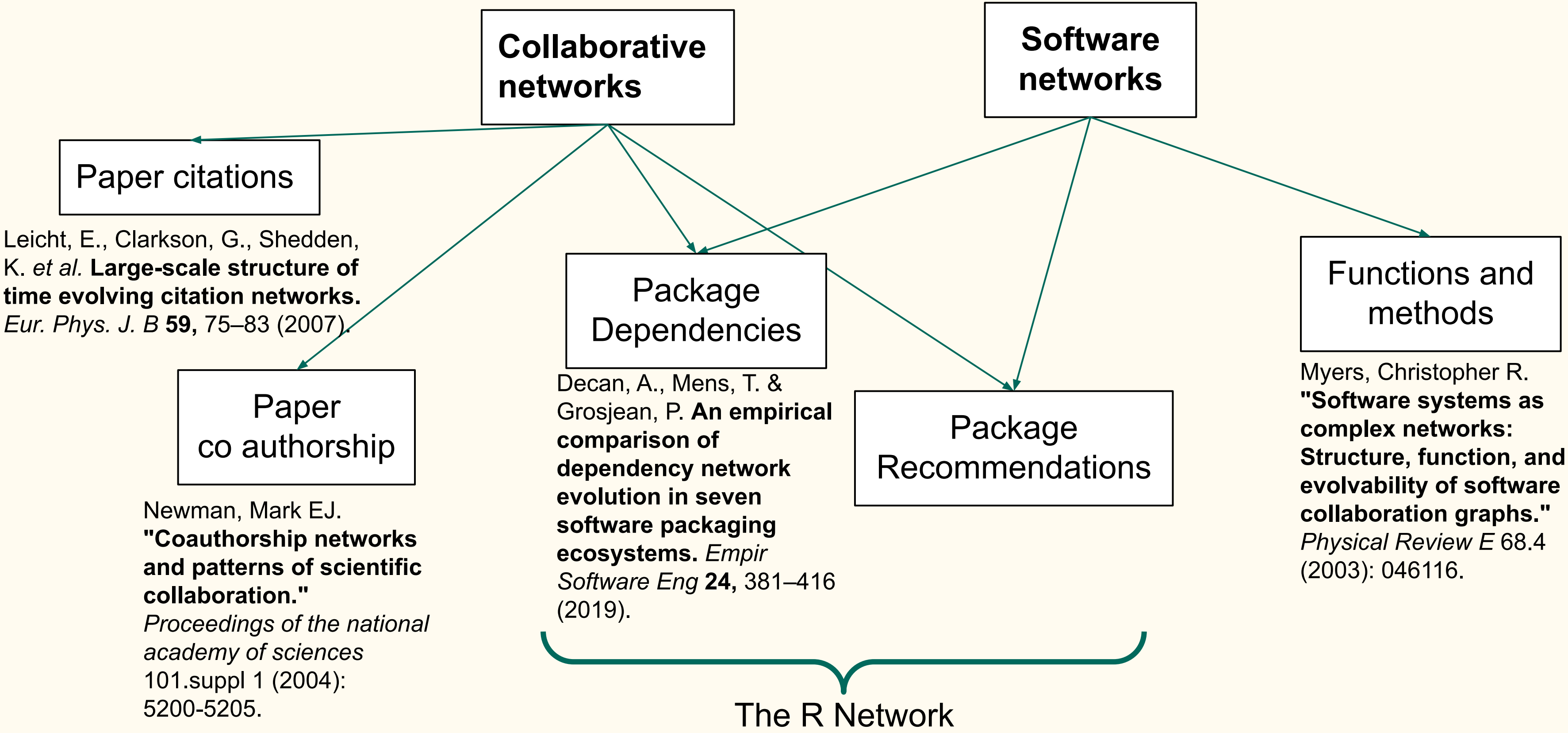
The R network evolution: Characterization of a collaborative software

Ariel Salgado - Inés Caridi

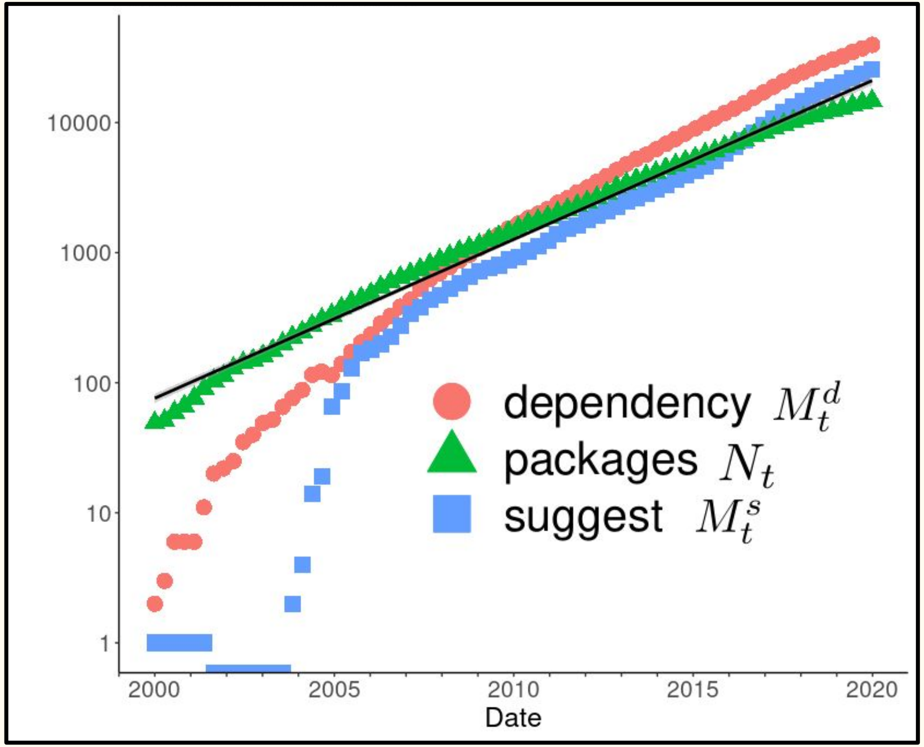
Instituto de Cálculo, FCEN - UBA, CONICET

WORKSHOP ON SOCIOPHYSICS: SOCIAL PHENOMENA FROM A PHYSICS PERSPECTIVE

Why software networks?



Why CRAN?

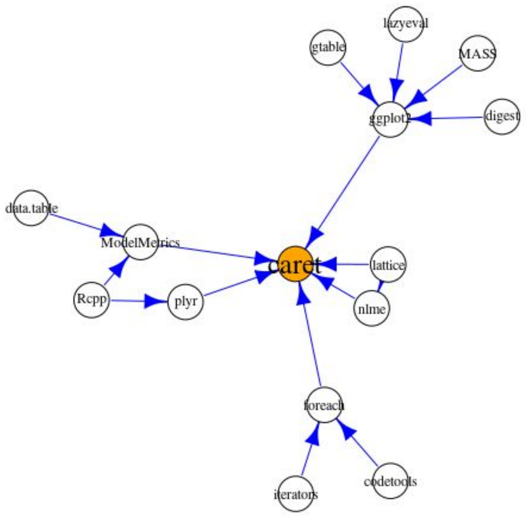


Packages + Dependencies + Suggestions

- Packages increased from less than 100 to more than 12 thousand in 20 years ($\boldsymbol{\log N_t \sim 7.7 \cdot 10^{-4} \cdot t}$)
- R started as a *niche* statistical language, while today is one of the preferred tools for Data Science.
- The growth of CRAN accompanies the growth of a worldwide community of users and developers.
- The network started being *sparse* but today the number of relations (**Dependencies and Suggestions**) surpasses the number of packages.

The comprehensive R Archive Network

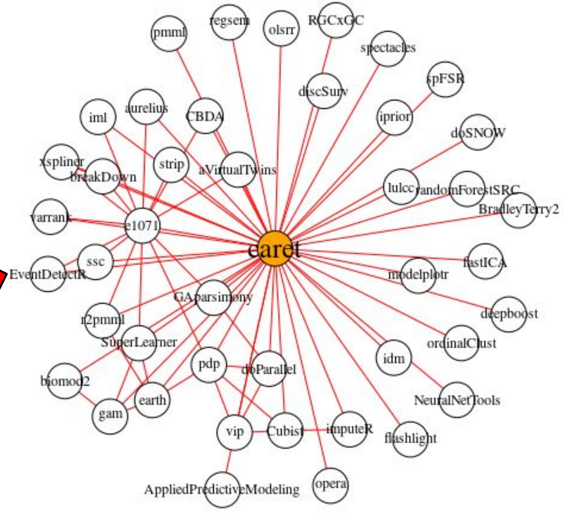
caret **dependency tree**



CRAN is represented through two networks:

- **Dependency network:** two packages are connected if one relies on the other to work.
- **Suggestion network:** two packages are connected if there is a tutorial if one package uses another in a tutorial.

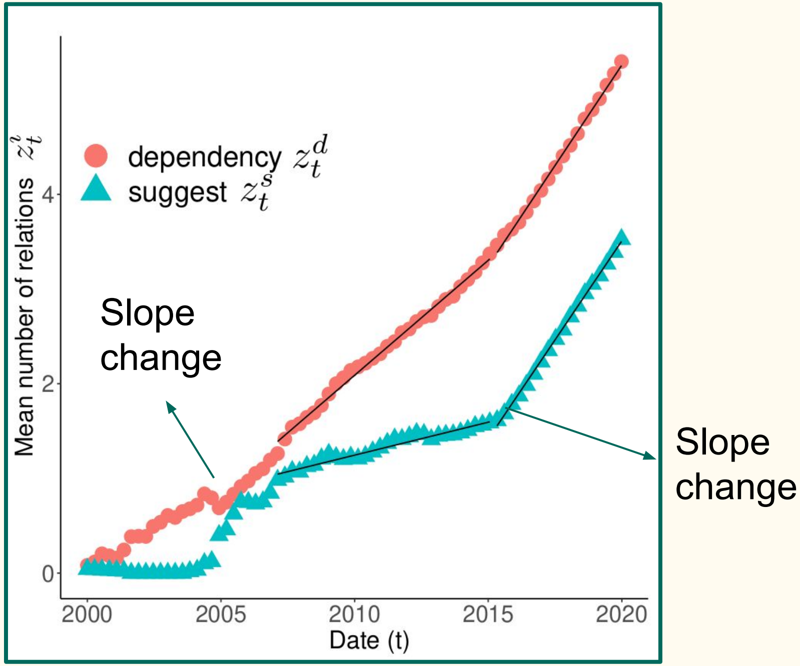
caret **suggestion neighbors**



In this talk...

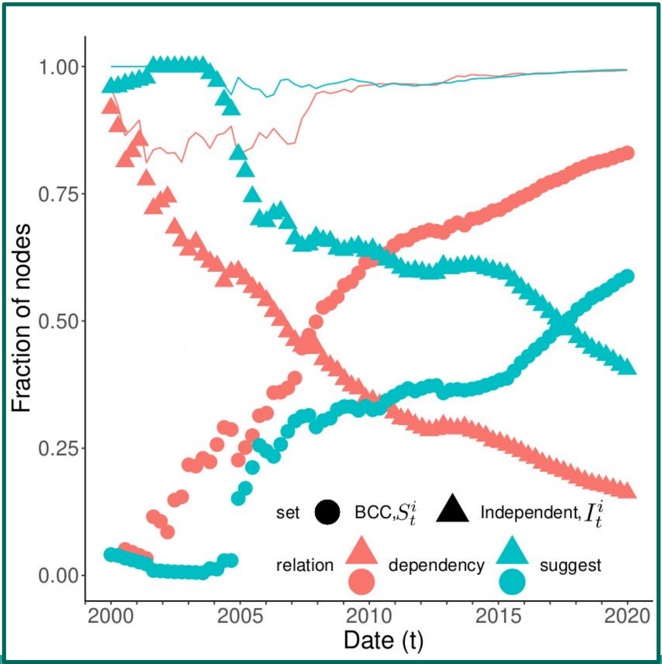
- Macroscopic growth of the network:
  - → Biggest connected component
  - → Mean degree
- Microscopic growth of the network:
  - → Degree distribution
  - → Connections at arrival
  - → Preferential attachment, and
- Commentary on the relationship between the network's events and the R events

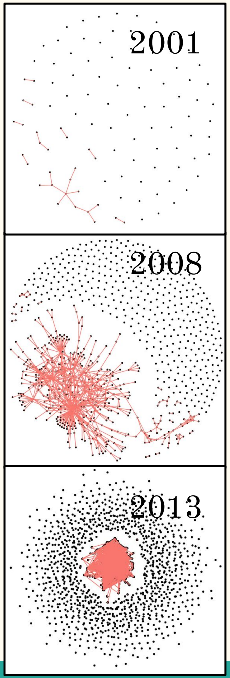
Macroscopic growth: mean degree and BCC



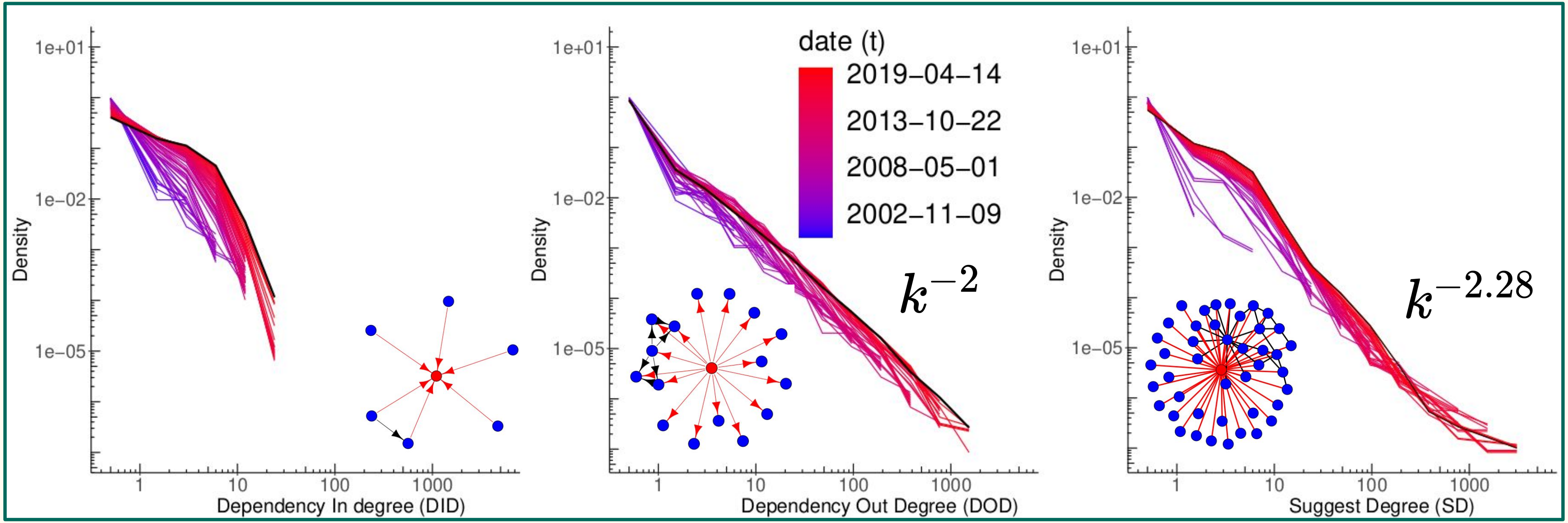
→ The mean degree changes its slope many times, indicating changes in the global connectivity, and probably in the developing logic

→Both networks transition from fully disconnected networks to mostly BCC.
→The structure is a balance between disconnected packages and the BCC





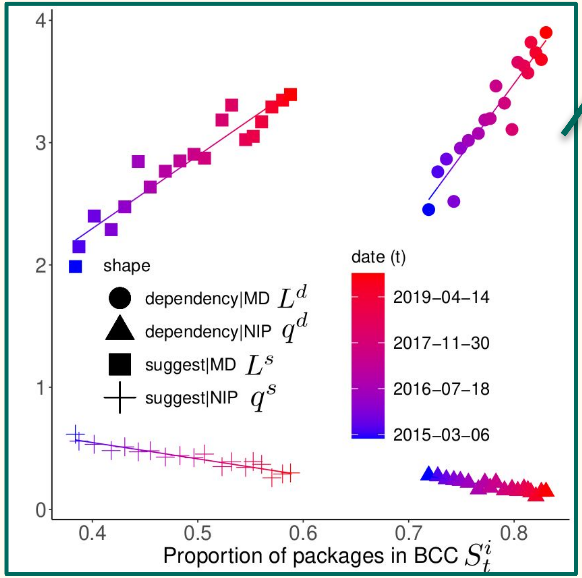
Microscopic view: degree distributions

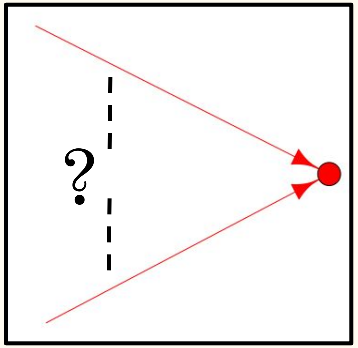


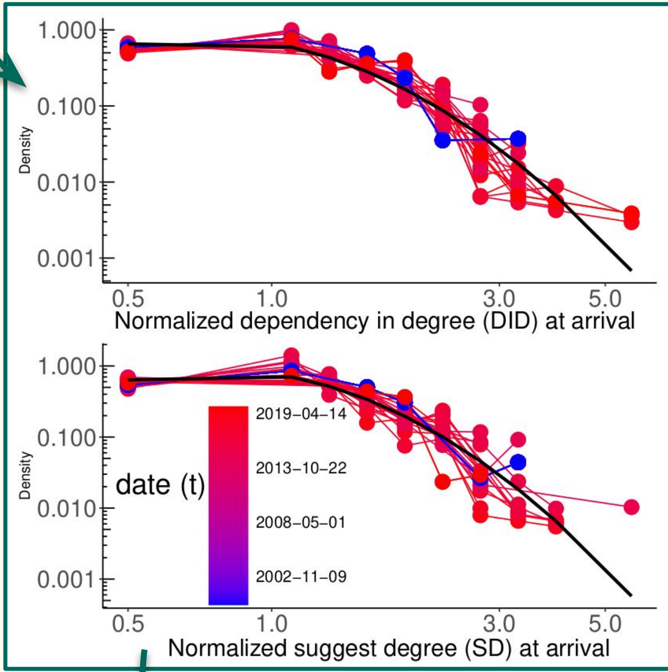
→The number of dependencies is bounded and resembles a lognormal distribution.
→Transition from a power law to a lognormal

→The number of suggestions and dependent packages resembles a power law.
→It does not change very much through evolution

Microscopic behavior: incoming degree distribution





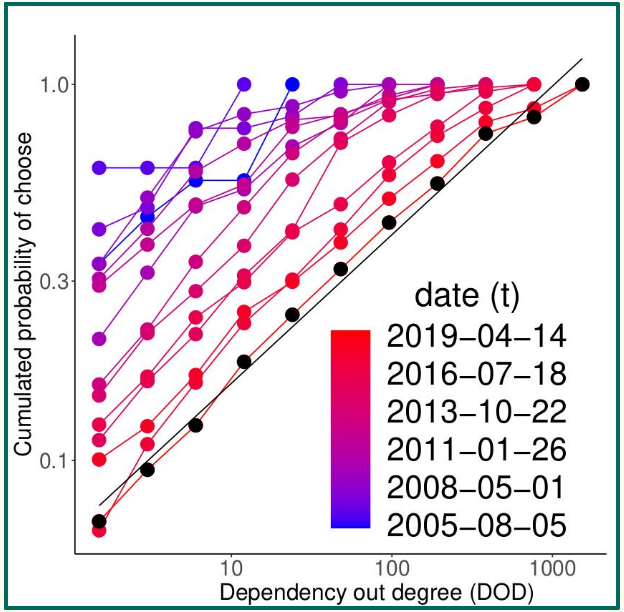


→ The distribution is a **zero inflated lognormal with mean scaled by the BCC**

→ The **number of connections** included by a new package **increases** as the fraction of packages in the **BCC increases**

$$P(k) = \begin{cases} a_0 S + b_0, \ k = 0 \\ \log \mathcal{N}\left(\frac{k}{a_1 S + b_1}\right), \ k > 0 \end{cases}$$

Microscopic behavior: preferential attachment



$$\Pi(k) \propto k^{-0.40} \rightarrow PA \propto k^{1.60}$$

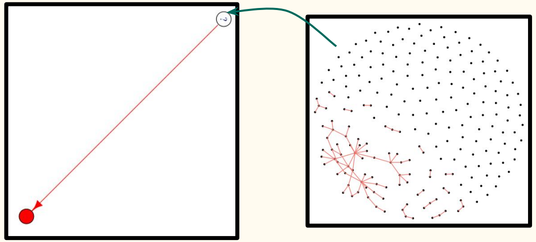
Following method in [1] we can visualize how **preferential attachment (PA)** changes through the evolution.

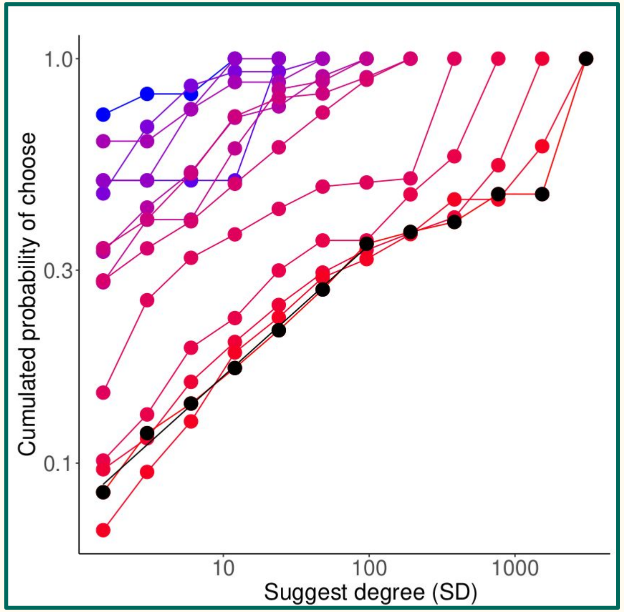
→ **Dependencies** show a power law PA.

→ **Suggestions** have near power law PA, including extra logarithmic terms

→ Both networks show evidence of **superlinear PA**

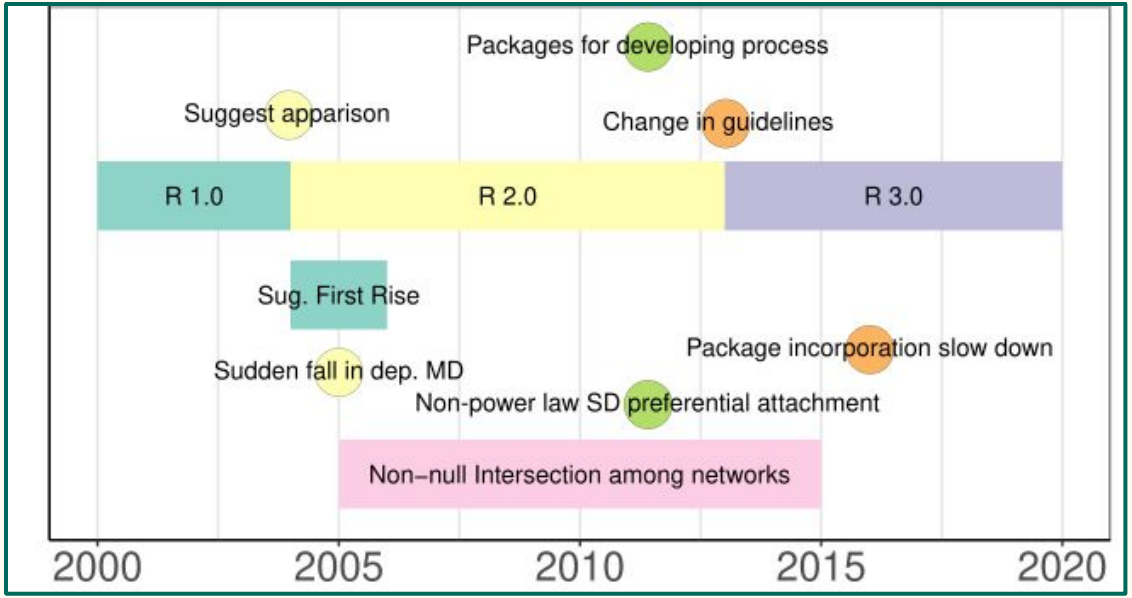
[1] H. Jeong, Z. Néda, A.-L. Barabási, Measuring preferential attachment in evolving network (2003)





$$\Pi(k) \propto k^{-0.32} \rightarrow PA \propto k^{1.96}$$

Sum up: Relation with historical events



→ Changes in versions of R produce changes in CRAN

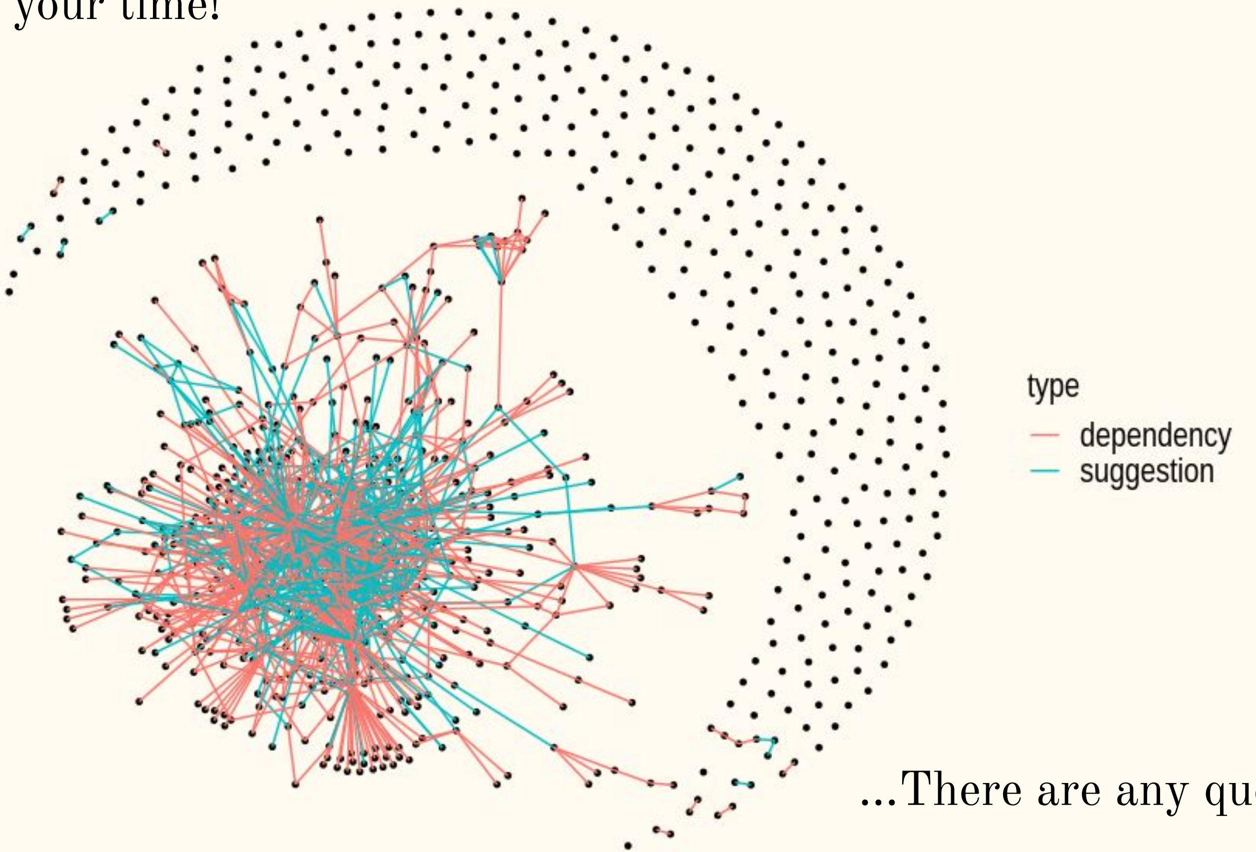
→ The suggestion PA changes due to the publication of packages aiding the development process.

→ The slow down in the number of packages can be due to a hardening of CRAN Publishing requirements

Conclusions

- **CRAN** is an example of an **empirical collaborative evolving network**,
- **External events** can be related to **growing patterns** and **connectivity changes**.
- **Dependency and suggestion** network **show preferential attachment**.
  Both are **superlinear.**
- A package tends to **require more packages as the BCC grows**. However, a **steady shape of the distribution remains**.
- Both networks can be seen as **one giant cluster** and **a myriad of independent packages**. As the network grows, the fraction of independent packages reduce and the giant cluster represents the biggest part of the network.

Thanks a lot for your time!



...There are any questions?