Information diffusion on realistic networks

Jean-Philippe Cointet[1†] and Camille Roth[1,2‡]

[1] *CREA, CNRS/Ecole Polytechnique, 1 rue Descartes, F-75005 Paris*
[2] *CRESS, Department of Sociology, University of Surrey, Guildford GU2 7XH, United Kingdom*

Les modèles de diffusion d'information mettent traditionnellement en jeu un réseau sous-jacent dont la topologie reproduit certaines propriétés observées dans les réseaux réels. Toutefois, la comparaison des phénomènes de diffusion observés sur des réseaux générés par des modèles classiques avec ceux se produisant au sein de réseaux réels reste peu étudiée. Dans une démarche empiriste, nous proposons dans cette étude d'évaluer l'écart de comportement induit par l'utilisation de divers modèles stylisés, dont notamment certains réseaux dits "sans-échelle".

Keywords: diffusion d'information, modèles de réseaux, validation empirique, simulation, structure de communautés.

Context and rationale

Although the issue of information diffusion has usually been appraised from a theoretical viewpoint, it stands, above all, for an empirical challenge: real-world networks, yet, are often represented by stylized hypotheses and it remains unclear how much diffusion dynamics may be affected by such assumptions. In particular, while information diffusion had been principally addressed by social scientists with a qualitative, ethnographic approach [Rog95, Val95], it progressively became crucial to correctly understand the effect of underlying network structures. This happened initially on stylized network structures using common network morphogenesis models such as Erdős-Rényi random graphs or simpler settings such as complete graphs or grid-based networks [ER59, Mor00, ZCB96].

These models seemed rather less realistic after a more empirical insight has recently been encouraged by findings from computer science and statistical physics concerning several kinds of real-world networks and highlighting a series of topological features, such as power-law ("scale-free") degree distributions, high clustering, short network diameter [WS98, FFF99, BA99, BT02, *inter alia*, in both Internet and social networks]. In particular, in the wake of a key result by [PSV01] suggesting that such networks have epidemiologic properties which are radically distinct from those of ER networks, in more recent diffusion models, scale-free networks are often used [GMT05, CAB06].

However, even if some authors insisted on the need for realistic topologies [ZCB96, BT02], it is unclear how accurately present models and corresponding analytic solutions or simulations render real-world phenomena. If indeed topology has a key impact on these phenomena, to which extent does the use of more or less stylized assumptions affect modeling results? On the other hand, to our knowledge, information diffusion models have scarcely been effectively simulated on real communication networks, while such approach obviously guarantees that all structural features of a real-world network are present.

In line with recent efforts at appraising the role of various topologies [GMT05, CAB06] the aim of this paper is thus to push the comparison to more realistic benchmarks. We focus on a particular communication network, an empirical social network, as a preliminary case study. We introduce a simulation framework to evaluate the accuracy of classical topologies in rendering phenomena occuring on this real-world network, using notably various stylized scale-free networks as (best) approximations of the real-world network.

[†] cointet@poly.polytechnique.fr
[‡] camille.roth@polytechnique.edu

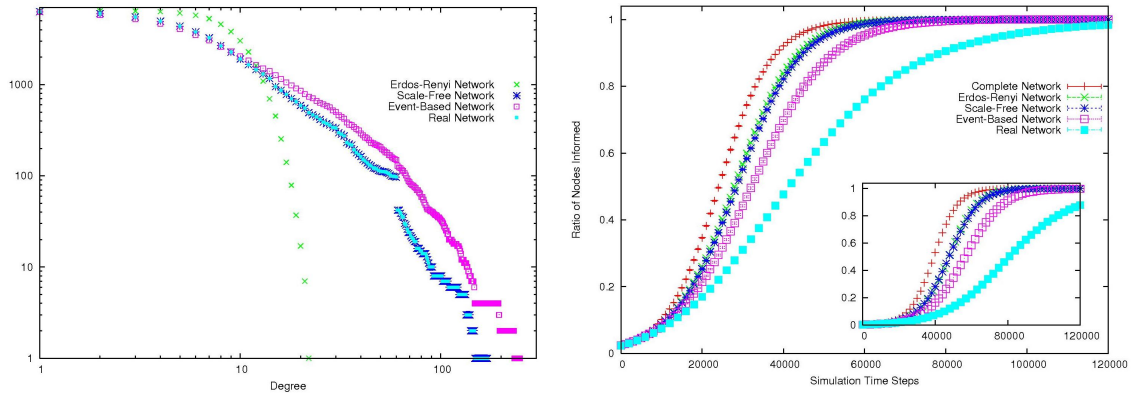

Figure 1: *Left:* cumulated degree distributions for the various network structures (*x-axis*: degree k, *y-axis*: $\mathcal{N}(k) = \sum_{k'=k}^{\infty} N(k')$). *Right:* Simulation results for complete, Erdös-Rényi, Scale-Free, Event-Based and real networks, using $\lambda = 0.02$ (outset) and $\lambda = 0.002$ (inset), along with associated 99% confidence intervals.

Simulation framework

We consider a set of N agents and a single piece of information: system state is described at any time t by a vector $c(t) \in \{0,1\}^N$, such that $c_i(t) = 1$ if the i-th agent knows the piece of information at t, otherwise $c_i(t) = 0$. We assume knowledge acquisition to be strictly growing: agents cannot forget. The initial setting is such that a given proportion λ of agents are initially "informed" (i.e. λN agents), while others are "ignorant" (i.e. $(1-\lambda)N$). Our discrete time simulation features knowledge exchange during interactions. At each time step an interaction occurs (between a target agent i, randomly chosen among the population, and one of its neighbors j, randomly selected), leading to transmission of information if the chosen neighbor j is informed ($c_j(t) = 1 \Rightarrow c_i(t) = 1$).

We measure the ratio of informed agents ρ among the whole population over time, $\rho(t) = \frac{1}{N}\sum_{i=1}^{N} c_i(t)$, while using a totally random initial spread of informed agents (only $\rho(0) = \lambda$). This part of the protocol remains extremely simple so that effects are both easily comparable across other conditions as well as already yielding significantly contrasted results.

Topologies

We consider a real network from the `Medline` database describing collaborations between embryologists working on the model-animal zebrafish: nodes are authors and links represent coauthorship. We consider only the largest connected component: 6453 agents, 67392 undirected links, from 2476 publications. Its degree distribution (Fig. 1) exhibits a power-law tail with a flatter head, typical of such social networks [BA99]. Starting from this real network (RN), we then investigate several network topologies created by progressively *degenerating* the original structure, i.e. by keeping less and less topogical features:

- *"Scale-Free" (SF)* network built from RN by reshuffling links while preserving the original degree distribution [MR95]. Total numbers of agents N and links M are also identical to RN, the original density $d = M/N(N-1)$ is thus preserved.
- *Erdös-Rényi (ER)* random graph [ER59] that only preserves the original RN density (same numbers of nodes N and links M), while the degree distribution is roughly Poisson.
- *Complete Network (CN)*, which only shares the number of agents N with RN — each agent is connected to all other agents, the total number of links is thus $N(N-1)/2$.

Clustering structure. Apart from degree distributions, the clustering structure is of particular interest; some authors suggest it might significantly alter information diffusion [BG98, Mor00]. The usual definition of "clustering" relies on the proportion of transitive triples, or average proportion of neighbors of node i who

are also connected together: $c_3(i) = \frac{[\text{number of pairs of connected neighbors of } i]}{k_i \cdot (k_i - 1)/2}$ with k_i degree of node i. Empirical networks are known to exhibit an abnormally high average clustering coefficient $\langle c_3 \rangle$, compared to those found in SF and ER random networks [BT02, NP03]; models traditionally endeavor to rebuild this statistical parameter as well. Here, RN also exhibits a high $\langle c_3 \rangle$ of .827 while SF has only a $\langle c_3 \rangle$ of .00539.

In an attempt to reconstruct a network topology that preserves both degree distribution and high $\langle c_3 \rangle$, we consider a last network model mimicking the original collaboration structure, i.e. its event-based structure:

- *Event-Based (EB):* we introduce a bipartite graph featuring agents on one side, events on the other side, and first assign to each agent and respectively to each event a "degree" drawn from empirical "agent-event" degree distributions. In other words, we thus maintain the original number of events featuring a certain number of agents, and reciprocally. Then, we link agents to events randomly, respecting their respective degrees. Finally, we compute the *projection of this graph* onto agents to build the collaboration network: two agents are linked when they participate in the same event.

EB network is closer to RN than SF in the sense that it keeps more topological features: it conserves the degree distribution, like in the SF case, it also preserves the $\langle c_3 \rangle$ clustering structure, unlike in the SF case (the clustering coefficient is high because of the clique-addition process precisely due to the projection) [GL04]. This is fairly well confirmed empirically, as shown on Fig. 1 and Tab. 1.

If we consider a recently-introduced, simple but finer clustering parameter, "$\langle c_4 \rangle$", which relies on the proportion of transitive diamonds or average proportion of common neighbors among the neighbors of a node i [LGH05] — $c_4(i) = \frac{\sum_{i_1=1}^{k_i} \sum_{i_2=i_1+1}^{k_i} \kappa_{i_1,i_2}}{\sum_{i_1=1}^{k_i} \sum_{i_2=i_1+1}^{k_i} \left[(k_{i_1} - \kappa_{i_1,i_2})(k_{i_2} - \kappa_{i_1,i_2}) + \kappa_{i_1,i_2} \right]}$ where κ_{j_1,j_2} is the number of nodes which the j_1-th & j_2-th neighbors of i have in common (leaving out i) — we see that EB falls short of one order of magnitude for $\langle c_4 \rangle$, suggesting that even EB still misses part of the community structure.

To summarize, we considered five distinct topologies: (i) real network RN, (ii) event-based network EB, (iii) scale-free network SF, (iv) Erdős-Rényi random graph ER, (v) complete network CN.

	RN	**SF**	**ER**	**CN**	**EB**
N	6453				
M	$6.74 \cdot 10^4$			$2.08 \cdot 10^7$	$7.62 \cdot 10^4$
d	.00162			.5	.00183
degree dist.	power-law tail		Poisson	—	power-law tail
$\langle c_3 \rangle$	.827	.00539	.00199	1	.753
$\langle c_4 \rangle$	.284	.000444	.000261	1	.0443

Table 1: Main characteristics of the various network structures: number of nodes N, number of links M, density d, degree distribution shape, clustering coefficients $\langle c_3 \rangle$ & $\langle c_4 \rangle$ (averages over 1000 networks for SF, ER & EB).

Simulation results

A set of $1,000$ random instances of each kind of network is created (excepted for real and complete networks, naturally unique). For each instance, we only work with the largest connected component, which is never made of less than 99.9% of the nodes (i.e. in the worst case, a negligible number of nodes is disconnected and left out). The simulation is initialized with a fraction λ of informed nodes. The asymptotic state of the system is obviously $\rho(\infty) = 1$.

On Fig. 1 is plotted the temporal evolution of the average ρ for each network topology. The closer we are from the real network, the slower the dynamics — CN performs the fastest, RN the slowest. More precisely and most surprisingly, ER and SF networks seem to behave *identically* with extremely similar convergence shapes. On the contrary, the behavior of EB is slower than other topologies, even being the best approximation of RN, yet still very unsatisfactorily rebuilding the original diffusion phenomenon.

Hierarchy in convergence speed is independent of λ (see Fig. 1): CN always performs fastest, followed by ER and SF, then EB, and finally RN, which is slower than on any other topology studied here.

It may thus well be far from sufficient to focus on the SF structure only: not all SF networks are equal [EK02, BT02] and, here, the most random SF network actually displays a behavior more similar to ER

than to RN. In contrast and in particular, EB results suggest a special influence of community structure in general, consistently with common claim in innovation studies that when agents are likely to interact more with agents they know and less with "remote" agents, it is less beneficial to knowledge propagation [Gra73, BG98] — denser clusters arguably provide more redundancies in the distribution of information among neighbors: as [Gra73] puts it, "if one tells a rumor to all his close friends, and they do likewise, many will hear the rumor a second and third time, since those linked by strong ties tend to share friends." A previous study by [BG98] argued that overlapping neighborhoods tend to make the diffusion of innovation slower, [EK02] later noted that epidemic threshold in highly clustered scale-free networks exhibited differences with the result for randomly-wired scale-free networks. This might explain why we get slower diffusion with EB than with SF, and even slower diffusion with RN, whose allegedly complex community structure could not be wholly reproduced, as illustrated by very different $\langle c_4 \rangle$ values.

Concluding remarks

Even with a very simple diffusion protocol, none of the topologies presented here rebuilds anything close to what the real network yields; here, SF behaves like ER but very differently from RN, opposite to common claim that SF topology is a crucial improvement over ER — local community structure seems to affect results even more than degree distributions, as partially suggests EB. The jury may thus still be out as to which topological features should absolutely be present in an artificial network for it to behave realistically for some given diffusion phenomena. Meanwhile, when possible, it could be safer to use real network topologies when developing diffusion models, through *simulations* which should first demonstrate that diffusion behaviors match.

References

[BA99] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[BG98] V. Bala and S. Goyal. Learning from neighbours. *Rev. Econ. Studies*, 65(3):595–621, 1998.

[BT02] T. Bu and D. Towsley. On distinguishing between internet power law topology generators. In *21st IEEE INFOCOM 2002*, volume 2, pages 638–647, 2002.

[CAB06] P. Crépey, F. Alvarez, and M. Barthélemy. Epidemic variability in complex networks. *Physical Review E*, 73(4):046131, 2006.

[EK02] V. Eguiluz and K. Klemm. Epidemic threshold in structured scale-free networks. *PRL*, 89:108701, 2002.

[ER59] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *Computer Communication Review*, 29(4):251–262, 1999.

[GL04] JL Guillaume and M Latapy. Bipartite structure of all complex networks. *IPL*, 90:215–221, 2004.

[GMT05] A. Ganesh, L. Massoulié, and D. Towsley. The effect of network topology on the spread of epidemics. In *IEEE Infocom*, volume 2, pages 1455–1466, 2005.

[Gra73] M. Granovetter. The strength of weak ties. *Am. Journal of Sociology*, 78(6):1360–1380, 1973.

[LGH05] P. Lind, M. Gonzalez, and H. Herrmann. Cycles and clustering in bipartite networks. *Physical Review E*, 72:056127, 2005.

[Mor00] S. Morris. Contagion. *Review of Economic Studies*, 67(1):57–78, 2000.

[MR95] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 161(6):161–179, 1995.

[NP03] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(036122), 2003.

[PSV01] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *PRL*, 86(14):3200–3203, 2001.

[Rog95] E. Rogers. *Diffusion of Innovations*. The Free Press, New York, USA, 4th edition, 1995.

[Val95] T. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, 1995.

[WS98] D. Watts and S. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[ZCB96] E. Zegura, K. Calvert, and S. Bhattacharjee. How to model an internetwork. In *IEEE Infocom*, volume 2, pages 594–602, San Francisco, CA, 1996. IEEE.