Gossip processes and small-world networks

Gesine Reinert

Department of Statistics
University of Oxford
reinert@stats.ox.ac.uk

Heilbronn Workshop, Bristol, March 19, 2013

Outline

Joint work with Andrew Barbour (Melbourne)

Small world networks

Has it happened to you? You meet a complete stranger, but in conversation it turns out that you have a common acquaintance. This experience is so surprising, yet at the same time so familiar, that we have coined a phrase to describe it: "It's a small world".

In a nutshell:
Surprisingly short distances between vertices despite only moderate local clustering

Some applications

- metabolic networks
- protein-protein interaction networks
- spread of epidemics
- neural network of *C. elegans*
- social networks
- collaboration networks (Erdös numbers ...)
- Membership of management boards
- World Wide Web
- interbank loans

Questions

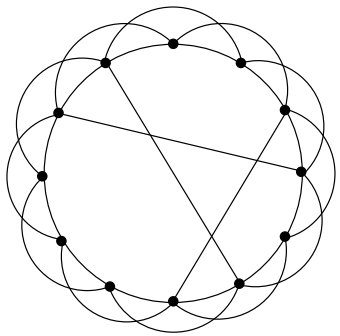
In this talk we address the following, related questions:

1. What is the shortest distance between two randomly chosen points?
2. If gossip spreads on a network, when has a certain proportion, say, 90 % of the population, heard the gossip? When has everyone heard the gossip?
3. If an epidemic spreads on a network, what is the proportion of susceptible individuals at any time point?

The Watts-Strogatz small world model

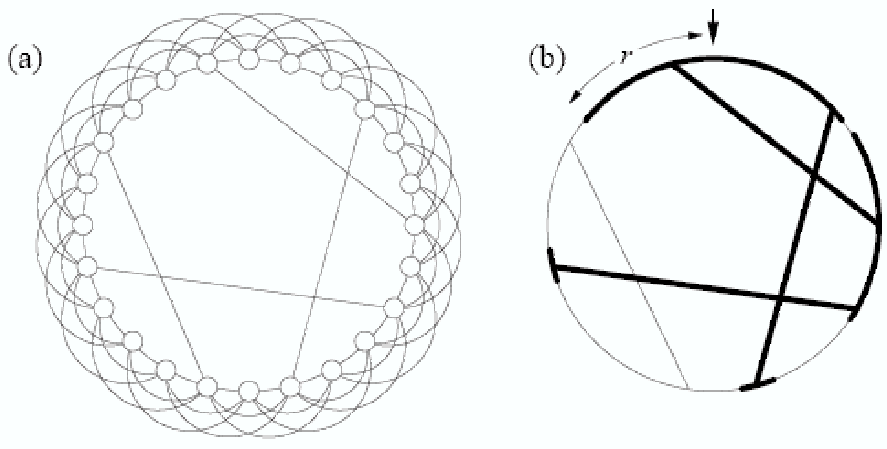
Milgram (1967): 6 degrees of separation?
Watts and Strogatz (1998); *Newman, Moore, Watts (2000)*; *Ball, Mollison, Scalia-Tomba (1997)*

A discrete circle with L vertices, and k nearest neighbours each, gets added random shortcuts, with ϕ probability of shortcut, per connection, so that there are $Lk\phi$ shortcuts on average.



Continuous approximation

The continuous (great) circle model (continuous small world model) on a circle C of circumference L has a Poisson $(L\rho/2)$ number of shortcuts added uniformly. Neighbourhoods are collapsed by dividing distances by k, and ρ corresponds to $2k\phi$. Chords between points have length zero.



A higher-dimensional version

Barbour and R. (2001)

Take $\mathrm{Po}\,(L\rho/2)$ shortcuts between random pairs of points in a finite, homogeneous space C in d dimensions, such as a sphere or a torus.

Local neighbourhoods are of the form $K(P, s)$, a ball of radius s around P; we assume that as $s \longrightarrow 0$,

$$|K(P, s)| \sim s^d v(K)$$

(and some more precise assumptions on this growth).

Configuration model

In a finite population of N vertices, N_k are of type k, that is, they have exactly k acquaintances.

An approximate construction is as follows: an individual of type k is equipped with k half-edges. Join all the half-edges into edges by a random matching of the half-edges.

The shortest distance in Watts-Strogatz small worlds

We use the continuous small world model, with a Poisson $(L\rho/2)$ number of shortcuts added uniformly to a circle C of circumference L. Assume $L\rho > 1$. Let $\mathcal{D}$ be length of the shortest path between two randomly chosen points

Barbour and R. (2001): uniformly in $|x| \leq \frac{1}{4}\log(L\rho)$,

$$P\left(\mathcal{D} > \frac{1}{\rho}\left(\frac{1}{2}\log(L\rho) + x\right)\right) \to \int_0^\infty \frac{e^{-y}}{1 + e^{2x}y}\, dy$$

as $L\rho \to \infty$ (and exact expression for bound on the distance).

The mean of the shortest distance is

$$\frac{1}{2\rho}\left(\log(L\rho) + \gamma\right),$$

where $\gamma \approx 0.577$ is the Euler constant.

Idea of Proof: Yule processes on the network

Pick a point P at random from C.

Our process walks from P at the same speed 2ρ in all possible directions, taking any shortcut that it can find.

Let $R(t)$ be the set of points that can be reached from P in time t.

Taking shortcut means initially creating a new intervals on the circle, but will in due time meet some areas that it has covered before; hence there is dependence in the temporal evolution.

Compare this process to a (Yule) pure growth process $S(t)$: start at P, with growth rate $2p$, ignoring overlap.

For small times t we expect $R(t) \approx S(t)$.

Pick another point P' at random from C, let an independent pure growth process run from that point. The time at which the two independent pure growth processes will meet will be $\approx \frac{1}{2}\mathcal{D}$.

More precisely...

For the pure growth process $S(t)$ started at P, let $M(t)$ be the number of intervals at time t and let $s(t)$ be the total length of the circle covered at time t. This process gives a Yule process with birth rate 2ρ; it is well-known that

$$\begin{aligned} \mathbf{E}M(t) &= e^{2\rho t}, \\ \mathbf{E}s(t) &= \frac{1}{\rho}\left(e^{2\rho t}-1\right). \end{aligned}$$

Let $N(t)$, $u(t)$ be the corresponding quantities for the pure growth process started at the point P'.

The time scale for intersections

Run both pure growth processes from time 0.

At time t there are approximately $e^{4\rho t}$ pairs of intervals and each has approximately length $\frac{1}{\rho}$.

Let V_t be the number of intersecting pairs of intervals at time t, one from the process started at P, the other from P', then

$$V_t \approx \frac{2}{L\rho} e^{4\rho t}.$$

At the time scale of the first encounter $\tau_x = \frac{1}{2\rho}\left\{\frac{1}{2}\log(L\rho) + x\right\}$ we have, roughly, $V_{\tau_x} \approx 2e^{2x}$.

Mixed Poisson approximation for V_t

Given $M(\tau_x) = m, N(\tau_x) = n$, and all the lengths $s_1, \ldots, s_m$ and $u_1, \ldots, u_n$ of the intervals,

$$V_{\tau_x} \approx \text{ Poisson} \left(\frac{2}{L} \sum_{i=1}^{m} \sum_{j=1}^{n} \min(s_i, u_j) \right).$$

Now

$$\{V_{\tau_x} = 0\} \approx \{\hat{V}_{\tau_x} = 0\} = \{\mathcal{D} > 2\tau_x\},$$

and thus

$$\begin{aligned} \mathbf{P}\{\mathcal{D} > 2\tau_x\} &\approx \mathbf{E} e^{-\frac{2}{L} \sum_{i=1}^{M(\tau_x)} \sum_{j=1}^{N(\tau_x)} \min(s_i, u_j)} \\ &= \mathbf{E} e^{-\frac{4}{L} \int_0^{\tau_x} M(v) N(v) dv}. \end{aligned}$$

Martingale argument

$$e^{-2\rho t} M(t) \quad \longrightarrow \quad W \text{ a.s.}$$

where W is exponentially distributed with parameter 1. So with W, W' indept. exp(1)

$$\exp\left\{-\frac{4}{L}\int_0^{\tau_x} M(v)N(v)dv\right\} \quad \approx \quad \exp\{-e^{2x}WW'\}$$

and

$$\mathbf{E}\exp\{-e^{2x}WW'\} = \int_0^{\infty} \frac{e^{-y}}{1+e^{2x}y}\,dy$$

as claimed.

In higher-dimensions

Take $\mathrm{Po}\,(L\rho/2)$ shortcuts between random pairs of points in a finite, homogeneous space C in d dimensions, such as a sphere or a torus.

Local neighbourhoods are of the form $K(P, s)$, which is ball of radius s around P. We assume that as $s \longrightarrow 0$,

$$|K(P, s)| \sim s^d v(K),$$

and that the probability of intersection of two subsets $K(P, t)$ and $K(Q, u)$ with P, Q uniformly chosen is approximately

$$L^{-1} v(K)(t + u)^d,$$

where L is the area of C.

The pure growth process

Again we approximate with pure growth processes with independently and uniformly positioned neighbourhoods. We define its neighbourhood size process by

$$\xi_t(A) = \#\{\text{neighbourhoods with radii having lengths in } A\}.$$

Let $M_0(t)$ be the number of neighbourhoods in the pure growth process at time t, and

$$M_l(t) := \int_{\mathbf{R}_+} x^l \xi_t(dx) = \sum_{j=1}^{M_0(t)} s_j^l, \quad l \geq 0,$$

be the sum of the l'th powers of the 'radii' of the neighbourhoods.

A martingale limit

Let

$$H_i(t) = M_i(t)\frac{\lambda_0^i}{i!},$$

and

$$\lambda_0 = \lambda_0(\rho) = (d!\rho v(K))^{\frac{1}{d}},$$

then, as $t \longrightarrow \infty$,

$$de^{-\lambda_0 t}H(t) \longrightarrow W_*(\infty)\mathbf{1} \text{ a.s..}$$

The shortest distance in d dimensions

Let D denote the distance between two randomly chosen points of C on the graph with a $\mathrm{Po}\,(L\rho/2)$–distributed random number of shortcuts. Then

$$\mathbf{P}\left[D > \frac{2}{\lambda_0}\left\{\frac{1}{2}\log(L\rho) + x\right\}\right] \to \mathbf{E}\left(\exp\left\{-de^{2x}W_*(\infty)W_*(\infty)'\right\}\right)$$

as $L\rho \to \infty$, uniformly in $|x| \leq \frac{1}{4}\log(L\rho)$, where $W_*(\infty)$ and $W_*(\infty)'$ are independent copies.

Beyond the shortest distance: gossip

The possibly simplest process is on a network is a *gossip* process: *Chatterjee & Durrett (2011), Aldous (2010)*

Information spreads locally from an individual to his neighbours on the two-dimensional torus, and also occasionally to other, randomly chosen members of the community.

Here we present our results from *Barbour and R. (2012)*.

Gossip and small world

In both models, a disc of informed individuals, centred on an initial informant, grows steadily in the torus; and long range transmissions of information occur in a Poisson process. Any such transmission contacts a randomly chosen point of the torus, initiating a new disc of informed individuals.

In a gossip process the rate of the Poisson process is proportional to the area (number) of informed individuals.

In *Barbour and R. (2001)* the Poisson process has rate proportional to the length of the boundary of the informed region.

The shortest distance for the gossip process

For the gossip process we get the same asymptotic behaviour as in the higher-dimensional small world with

d replaced by $d+1$ and

$\lambda_0(\rho)$ replaced by

$$\lambda_0 := (d!\rho v(K))^{\frac{1}{d+1}}.$$

We use the notation λ_0 throughout, with the two different definitions depending on the process.

Two observations I

In both processes, an inspection of the proofs shows that the branching process approximation holds beyond the initial phase. The mean number of self-intersections in the approximating branching process is small up until times t at which

$$e^{2\lambda_0 t} \asymp \Lambda := L\lambda_0^d / v(K);$$

i.e.

$$t = t_{\Lambda,x} = \frac{1}{2\lambda_0}\{\log \Lambda + x\}.$$

Up to this time the branching process approximation works well.

Two observations II

We could think of our argument for the shortest distance as follows. We start the process, call it Y_{P_0}, at a random point P_0 forwards in time. We pick another point P at random and run the process, call it $\bar{Y}_P$, backwards in time: we keep track on who would inform P about the gossip if they know the gossip themselves.
Then in the gossip model $Y_{P_0}(t)$ and $\bar{Y}_P(t, 2t)$ are independent.

What happens next?

Let $L^{-1}V_{P_0}(t)$ be the proportion of the volume of C that is covered at time t (starting the process in P_0).

We shall see:
Once the initial stages have passed and we are on the time scale for shortest paths to occur, $L^{-1}V_{P_0}(t)$ grows more or less deterministically. More precisely, if $\mathcal{F}_s$ denotes the $\sigma-$algebra with all information up to time s then we can show that for $t > t_{\Lambda,x}$ with $|x| \leq \frac{1}{2}\log\Lambda$,

$$L^{-1}V_{P_0}(2t) \approx L^{-1}\mathbf{E}\{V_{P_0}(2t)|\mathcal{F}_s\}.$$

Heuristics

Note that

$$L^{-1}\mathbf{E}V_{P_0}(2t) = \mathbf{P}_{P_0}[d_{SW}(P_0, P) \leq 2t] = \mathbf{P}[Y_{P_0}(t) \cap Y_P(t) \neq \emptyset].$$

For $t > s$,

$$\begin{aligned} & L^{-2}\mathbf{E}\{[V_{P_0}(2t)]^2 \,|\, \mathcal{F}_s\} \\ &= \mathbf{P}_{P_0}[\{d_{SW}(P_0, P) \leq 2t\} \cap^* \{d_{SW}(P_0, P') \leq 2t\} \,|\, \mathcal{F}_s] \\ &\sim \mathbf{P}[\{Y^*_{P_0}(t) \cap^* Y^*_P(t) \neq \emptyset\} \cap \{Y^*_{P_0}(t) \cap^* Y^*_{P'}(t) \neq \emptyset\} \,|\, \mathcal{F}^*_s] \\ &\sim \{\mathbf{P}[Y^*_{P_0}(t) \cap^* Y^*_P(t) \neq \emptyset \,|\, \mathcal{F}^{P_0}_s]\}^2 \end{aligned}$$

where $Y^*_{P_0}$, Y^*_P and $Y^*_{P'}$ come from three independent branching processes, and $\mathcal{F}^*_s$ denotes the history of the branching process for $Y^*_{P_0}$ up to s.

The limiting behaviour

To describe the limiting behaviour we use the Laplace transform of $W_*(\infty)$,

$$\varphi(\theta) := \mathbf{E}\{e^{-\theta W_*(\infty)} \mid M_0(0) = 1, M_j(0) = 0, j \geq 1\},$$

and define

$$h(t) = h_d(t) = 1 - \varphi(e^t).$$

The spread of gossip: Theorem

There exists a random variable U such that

$$\mathbf{P}\left[\sup_x \left| L^{-1} V_{P_0}\left(\frac{\log \Lambda + x}{\lambda_0}\right) - h_d(x + \log C_d + U)\right| > 4\Lambda^{-a_1}\right] \leq c\Lambda^{-a_2}$$

for some $a_1, a_2 > 0$ and $c < \infty$. Here e^U has the distribution of $W_*(\infty)$ and $C_d = \frac{d!}{d+1}$.

Observations

h_d depends on C only through the dimension;
λ_0 is the initial exponential growth rate for the branching process;
U can be viewed as a random delay, caused by early fluctuations in the growth of the process, before the deterministic evolution governed by h_d sets in.

Complete coverage

Assume that, for each s, C can be covered by $n(s)$ islands of the form $K(P, s)$, where $n(s)$ satisfies

$$n(s) \;\leq\; c_0 L/\{v(K)s^d\}, \qquad 0 < s < L^{1/d}.$$

for some c_0.

Then, except on a set of probability of order $O(\Lambda^{-\delta})$, for some $\delta > 0$, the whole of C is covered before time

$$\tau(\Lambda, s) + \frac{2}{\lambda_0}\left\{\frac{72\log\Lambda}{d!}\right\}^{1/d},$$

where $\tau(\Lambda, s) = \lambda_0^{-1}\{\log\Lambda + O((\log\Lambda)^{1/(d+1)})\}$.

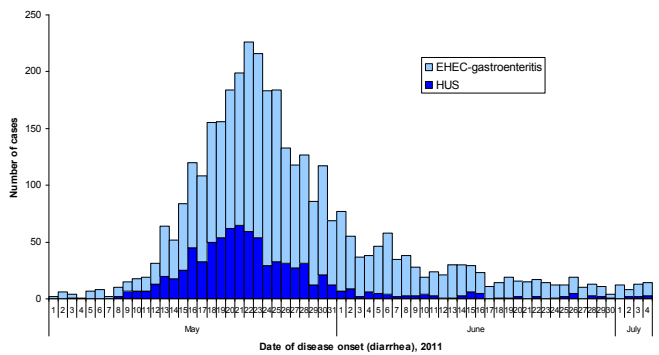
Aldous (2010) conjectured and Chatterjee & Durrett (2011) proved a limit law on the torus, with the limit implicitly described with help of a function h which satiesfies a certain integral equation. They give a rough error bound.

Here we identify the function h as a Laplace transform, and we give an estimate of the approximation error that is uniform for all time.

An S-I-R epidemic on a configuration graph

In a finite population of size N, we suppose that pairs of individuals are either acquainted with one another or are not, and that infectious contacts can only be made between graph neighbours.

Example for an epidemic curve: E.coli outbreak 2011 (*Werber et al. 2012*)



Assumptions

N_k members of the population are 'type k' individuals with exactly k acquaintances, $\sum_{k=1}^{K} N_k = N$; $N_k \in \{\lfloor Np_k \rfloor, \lceil Np_k \rceil\}$, for fixed $p_1, \ldots, p_K$, and $M := \sum_{k=1}^{K} kN_k$ is even;
there is a finite, N-independent upper bound K on the number of acquaintances that an individual may have.

An infected type k individual makes contact with a given type l acquaintance at a random time after infection that has distribution function G_{kl} and is independent of all other contact times.

A type k individual remains infectious for a random time with (possibly defective) distribution Φ_k, again independently of everything else.

Construction

Start with one initially infected individual. Match its half–edges by random choice from the set of all half–edges; then attach the infectious period to the initial individual, and the lengths of time to potentially infectious contact to the edges.

This yields a set of infected vertices, together with the times of their infection, some of which may be infinite.

Repeat the process for the first of these vertices (if any) to be infected.

Proceed in this way, always choosing for development the infected vertex with unmatched half–edges that has the smallest time of infection, until the first time that either at least $\lfloor \sqrt{N} \rfloor$ vertices have been infected or the infection dies out.

If the infection does not die out, there remains a set of infected vertices whose subsequent contact history has not been explored.

If a half–edge is picked for a second or subsequent time, ignore the choice and re-sample until a new one is chosen; if a vertex is chosen that has already been infected, ignore it for future development.

The process viewed backwards

For the process seen backwards from a randomly chosen individual, carry out essentially the same procedure for a specified time; but the vertex to which the infectious period is attached, is that of the child, not the parent.

Half–edges that have previously been used, including those that were used in the forward process, are discarded and re-sampled; the half–edges that are associated with the set of infected but unexplored vertices from the forward phase are still available for choice, and are those that close chains of infection.

The forward branching process approximation

We can approximate the infection process by an age–dependent multitype branching process with K types.

A type k individual (other than the initial individual) has $k-1$ offspring and each of these is of type l with probability lp_l/m, where $m = \sum_{l'=1}^{K} l' p_{l'}$.

The type k individual also has an infectious period randomly assigned to it from the distribution Φ_k, and the times to contact along the different edges are assigned independently from the appropriate distributions G_{kl}.

The times of birth of the descendants of a given individual are dependent, because they are finite only if they do not exceed the infectious period of the common parent.

The backward branching process approximation

The offspring distribution is identical, but the infection times of the offspring of a given individual, although having the same marginal distributions as before, are now independent, because the relevant infectious period, determining whether a contact results in infection, is that of the child, and not of the parent.

Choosing the contact times for type k – type l contacts *independently* from G_{kl} and the infectious periods independently from the Φ_k means that the times to birth in the backward branching process have distributions that do not depend on N.

Branching process quantities

Define a matrix μ by

$$\begin{aligned}\mu_{lk}(s) &= (l-1)\{kp_k/m\}\int_0^\infty e^{-su}(1-\Phi_l(u))G_{lk}(du)\\ &=: (l-1)\{kp_k/m\}U_{lk}(s),\end{aligned}$$

and write $\mu_{lk} := \mu_{lk}(0)$. Let $\mu^{(1)}(s)$ be obtained from $\mu(s)$ by removing the first row and column.

Assume that the matrix $\mu^{(1)}(0)$ has dominant eigenvalue larger than 1, and define the *Malthusian parameter* λ to be such that $\mu^{(1)}(\lambda)$ has dominant eigenvalue equal to 1.

Then let ζ^T be the left eigenvector of $\mu^{(1)}(\lambda)$ with eigenvalue 1 so that $\zeta^T 1 = 1$.

The backwards process

We now have

$$\hat{\mu}_{lk}(s) \;:=\; (l-1)\{kp_k/m\}U_{kl}(s),$$

and the Malthusian parameter is still λ. The left eigenvector of $\hat{\mu}(\lambda)$ for eigenvalue 1 is given by $\hat{\zeta}^T$, scaled so that $\hat{\zeta}^T 1 = 1$. Let $\mathbf{W}_*^{(i)}$ be the limiting random variable for the backward process starting with a single individual of type i;

$$\widehat{B}'(t)e^{-\lambda t} \;\to\; \mathbf{W}_*^{(i)}\hat{\zeta} \text{ a.s.}$$

Then $\mathbf{W}_*^{(i)}$ has a distribution whose Laplace transform $\hat{\psi}_*^{(i)}$ can be found from the solutions to a set of implicit equations belonging to the backward branching process whose individuals, including the initial individual, all follow the same rules.

Define

$$\tau_N := \inf\{t > 0 : \sum_{l=1}^{K} B_l'(t) \geq \lfloor \sqrt{N} \rfloor\}$$

and

$$m_*^{(2)} := \frac{1}{m} \sum_{k=1}^{K} \sum_{l=1}^{K} \zeta_k (k-1)(l-1) \hat{\zeta}_l \int_0^{\infty} \lambda v e^{-\lambda v} (1 - \Phi_k(v)) G_{kl}(dv) .$$

The epidemic curve: our main result

Suppose that the forward branching process is supercritical. Then the total proportion of susceptibles

$$N^{-1}\sum_{l=1}^{K} S_{Nl}(\tau_N + \lambda^{-1}\{\tfrac{1}{2}\log N + u\})$$

is well approximated by

$$\sum_{l=1}^{K} p_l \hat{s}_l(u),$$

uniformly in u.

Here $\hat{s}_l$ is the decreasing function given by

$$\hat{s}_l(u) \;:=\; \hat{\psi}_*^{(l)}(e^u m_*^{(2)}).$$

(Barbour + R., 2013)

A special case: the Volz equations

Volz (2008) assumes $\Phi(v) = 1 - e^{-\beta v}$ and $G(dv) = \alpha e^{-\alpha v} dv$. This gives

$$\begin{aligned} h(u) &= 1 - \int_{(0,\infty)} \{1 - m^{-1} g'(h(u-v))\} \alpha e^{-(\alpha+\beta)v} \, dv \\ &= \frac{\beta}{\alpha+\beta} + \frac{1}{m} \int_{-\infty}^{u} g'(h(w)) \alpha e^{-(\alpha+\beta)(u-w)} \, dw \,, \end{aligned}$$

where $g(s) = \sum_{k=1}^{K} p_k s^k$.

Differentiate:

$$\frac{dh}{dt} = \frac{\alpha}{m} g'(h) - (\alpha+\beta)h + \beta = (\alpha+\beta)(\tilde{f}(h) - h),$$

where $\tilde{f}(s) = (\alpha g'(s) + \beta)/(\alpha+\beta)$, which is a probability-generating function.

Hence $h(\infty)$ is the solution $\tilde{q}$ smaller than 1 to the equation $\tilde{f}(s) = s$, and the asymptotic final proportion of susceptible individuals at the end of a large outbreak is given by $g(\tilde{q})$.

Universality?

For the shortest path length, *Bhamidi, van der Hofstad and Hooghiemstra (2012)* showed a similar result, using similar ideas, for the configuration model, and for a range of related models, with weights, but without bounds. *Barbour and R. (2011)* showed a similar result, using similar ideas, for random multitype intersection graphs.

The behaviour may be typical for small world networks as long as the branching process approximations hold.

Other networks?

An open problem is whether the spread of gossip, and epidemics, on other networks display similar behaviour. In the gossip proofs and in the epidemics proofs, rather than first constructing a network and then running a process on it, we construct the network jointly with the process on the network. This approach may be promising for other network types.