A fixed point approximation for a routing model in equilibrium

Malwina Luczak

School of Mathematical Sciences
Queen Mary University of London
UK
e-mail: m.luczak@qmul.ac.uk

Brisbane, July 2013

Complex random networks

- Many complex systems (e.g. internet, biological networks, communications and queueing networks) can be modelled by Markov chains.
- Under suitable conditions there is a law of large numbers, i.e. the random system can be approximated by a deterministic process with simpler dynamics, derived from average drift.
- One may want to establish such a law of large numbers, with quantitative concentration of measure estimates, in a non-stationary (time-dependent) regime, or in equilibrium.

Load-balancing

- The most basic load-balancing model is as follows. There are n bins, and we throw n balls sequentially. At each step, the current ball examines d ($d \geq 1$) bins chosen uniformly at random with replacement, and is placed in one with smallest load. What is the maximum load of a bin at the end?
- When $d = 1$, then with high probability the maximum load of a bin is of the order $\log n / \log \log n$. When $d \geq 2$, then with high probability the maximum load of a bin is $\log \log n / \log d + O(1)$ (Azar et al., 1994).

Power of two choices

- This is called the power of two choices phenomenon, and has important implications for performance of networks.
- It means that by allowing calls/tasks to choose the best among an even very small number of alternatives we can dramatically improve the performance of the network, while still keeping routing costs (in terms of e.g. time taken to examine different options) low.
- This idea has now been around for over 20 years, and studied in the context of various models.

Communication network model with alternative routing

- We have a fully connected *communication graph* K_n, with vertex set $V_n = \{1, \ldots, n\}$ and edge set $E_n = \{\{u, v\} : 1 \leq u < v \leq n\}$.
- Each link has a capacity of $C = C(n)$ units ($C \in \mathbb{Z}^+$, bounded or $C \to \infty$ with n).
- $N = \binom{n}{2}$ calls arrive over N time steps, one at a time.
- Every new call chooses its endpoints uniformly at random, so that every edge $\{u, v\} \in E_n$ is chosen with probability $1/N$.

Communication network model with alternative routing

- If the link joining u and v has spare capacity (i.e. is currently carrying fewer than $C(n)$ calls), then we route a newly arriving call onto that link.
- Otherwise, we select d ($d \geq 1$) intermediate nodes $w_1, \ldots, w_d \in V_n \setminus \{u, v\}$ uniformly at random with replacement, and try to route the call along one of the two-link paths $\{u, w_i\}, \{v, w_i\}$ (an *alternative route*) for some $i = 1, \ldots, d$.

- How we choose among these d paths may depend in some (possibly complicated) manner on their current loads only. (We call such routing strategies General Dynamic Alternative Routing algorithms or GDAR algorithms.)
- If none of the d chosen paths has spare capacity (i.e. if on every one of them, at least one link has $C(n)$ calls in progress), then the new call is lost.
- Each successfully routed call occupies its path till the end of the process.

BDAR and FDAR

- Two particular types of GDAR's have been studied before.
- The First Dynamic Alternative Routing algorithm always chooses the first possible two-link route among the d chosen (i.e. first on the list of choices), if there is one among the d chosen where both links carry less than $C(n)$ calls at the time of the arrival.
- The Balanced Dynamic Alternative Routing algorithm chooses an alternative route which minimises the larger of the current loads on its two links, if possible. (Ties may be decided e.g. at random, or by selecting the first best route on the list.)

Dynamic version

- Calls arrive in a Poisson process at rate λN, where $\lambda > 0$ is a constant, and $N = \binom{n}{2}$.
- Each new call chooses its route as in the 'static' version.
- Accepted call durations are unit mean exponential random variables, independent of one another and of the arrivals and choices processes.
- Every call that is accepted into the system (on either a one-link or a two-link path) occupies one unit of capacity on each link of its route for its duration. When a call terminates, one unit of capacity on each link of its route is freed.

Difficulties with analysing the routing model

- Note that transitions may change the state of more than one link (when the call is routed on an alternative route).
- Thus, in comparison with the basic load-balancing model described earlier, this model has a lot less symmetry.
- For example, here the distribution of a pair of link loads may depend on whether or not they have a node in common.

Earlier results for the BDAR/FDAR routing model

The BDAR and FDAR algorithms (not under this name) were studied by L. and Upfal (1999), who first observed the following:

- For BDAR with $d \geq 2$, link capacity $C(n)$ of the order $\log \log n / \log d$ is sufficient to ensure that, in equilibrium, all calls arriving into the system during an interval (or all N calls in the static version) are routed successfully whp (ie with probability tending to 1 as $n \to \infty$).
- For FDAR, for any fixed d, link capacity has to be at least of the order $\sqrt{\log n / \log \log n}$ for this to be the case.

Earlier results for the BDAR/FDAR routing model

- More precise versions of these results can be found in L., McDiarmid and Upfal (2003) for the static version, and L. and McDiarmid (2013+) for the dynamic version.
- Also, the version with $d = 1$ and constant capacity C was studied from a very different perspective by Gibbens, Hunt and Kelly (1990), Crametz and Hunt (1991) and Graham and Meléard (1993).

With the exception of the work of Gibbens, Hunt and Kelly (1990), Crametz and Hunt (1991) and Graham and Meléard (1993), the papers mentioned above in fact do not analyse the model as described, but a (slightly easier) version, where the capacity of each link $\{u, v\}$ is split into three parts.

One part of each link ($C_1(n)$ units) is reserved solely for direct calls, and the others for calls on two-link paths, with one end u and with one end v respectively ($2C_2(n)$ units).

- Equivalently, for every pair $\{u, v\}$ of distinct nodes, there is a direct link, also denoted by $\{u, v\}$ with capacity $C_1(n)$. Also, there are two indirect links, denoted by uv and vu, each with capacity $C_2 = C_2(n)$.
- The indirect link uv may be used when for some w a call $\{u, w\}$ finds its direct link saturated, and we seek an alternative route via node v. Similarly, vu may be used for alternative routes for calls $\{v, w\}$ via u.
- Additionally, L. and McDiarmid (2013+) do not use direct one-link paths at all, but instead demand that each call be routed along a path consisting of a pair of indirect links.

FDAR algorithm

- We take $\lambda > 0$, d fixed. Also, $C = C(n) \sim \alpha \frac{\ln n}{\ln \ln n}$ as $n \to \infty$.
- 'Burn-in' period t_0: if the distribution of the initial state X_0 is stochastically dominated by the stationary distribution π, let $t_0 = 0$, and otherwise let $t_0 = t_0(n) = 5 \ln n$.
- Let $t_1 \geq t_0$, and consider intervals $[t_1, t_1 + n^K]$.

- α is K-*good* if, whatever version of GDAR we use, for each $t_1 \geq t_0$, the mean number of calls lost during the interval $[t_1, t_1 + n^K]$ is $o(1)$; and α is K-*bad* if, when we use FDAR, for each $t_1 \geq 0$, the mean number of calls lost during the interval $[t_1, t_1 + n^K]$ is $n^{\Omega(1)}$. (Observe that α cannot be both K-good and K-bad.)

- Theorem (L. and McDiarmid (2013+))

  If $\alpha > 2/d$ then α is K-good for some $K > 0$, and if $\alpha \leq 2/d$ then α is K-bad for each $K > 0$.

Theorem (L. and McDiarmid (2013+))

Let $\alpha > 2/d$ *and let* $K > 0$.

(a) If $2/d < \alpha \leq 1$ *(and so* $d \geq 3$*) then* α *is* K*-good for* $d\alpha - K > 2$*, and* α *is* K*-bad for* $d\alpha - K < 2$.

(b) If $\alpha \geq 1$ *(as must be the case when* d *is 1 or 2) then* α *is* K*-good for* $\alpha - K > 3 - d$*, and* α *is* K*-bad for* $\alpha - K < 3 - d$.

BDAR algorithm

Theorem (L. and McDiarmid (2013+))

Let $\lambda > 0$ be fixed and let $d \geq 2$ be a fixed integer. Let $K > 0$ be a constant. Then there exist constants $\kappa = \kappa(\lambda, d)$ and $c = c(\lambda, d, K) > 0$ such that the following holds.
(a) Suppose that $C(n) \geq \ln\ln n/\ln d + c$ and we use the BDAR algorithm. Let $t_0 = 0$ if X_0 is stochastically dominated by π, and let $t_0 = \kappa \ln n$ otherwise. Then the expected number of failing calls during $[t_1, t_1 + n^K]$ is $o(1)$ for each $t_1 \geq t_0$.
(b) If $C(n) \leq \ln\ln n/\ln d - c$ and we use any GDAR algorithm, then whp at least $n^{K+2-o(1)}$ calls are lost during the interval $[t_1, t_1 + n^K]$ for each $t_1 \geq 0$.

- Just like for other models of large networks, we would like more precise information about the distribution of link loads, in a transient regime and in equilibrium.
- We would also like to analyse the 'original' model, without splitting link capacities, in the case $d = 1$.
- In the case $d \geq 2$, the model without splitting link capacities does not exhibit the power of two choices phenomenon.

Other earlier work

In the case $d = 1$ and capacity C constant, a law of large numbers for this model is conjectured by Gibbens, Hunt and Kelly (1990) and shown by Crametz and Hunt (1991). (See also Graham and Meléard (1993).)

More precisely, under suitable initial conditions, for each constant time $t > 0$ and each $k \in \{0, \ldots, C\}$, the proportion of links in the network that have load k at time t is close to a deterministic function $\xi_t(k)$, where (ξ_t) solves a $(C+1)$-dimensional differential equation.

- All previous results have been non-quantitative, and restricted only to the special case of the model with $d = 1$ and capacity C constant.
- Also, Graham and Meléard's results work only for special kinds of initial conditions. In particular, they assume that initially all nodes are exactly exchangeable.
- Before our work, there have been no laws of large numbers results in equilibrium, and no results on the speed of convergence to equilibrium.

Questions

- Is the total number of links with load k ($k = 0, 1, \ldots, C(n)$) at a given time well-concentrated around its expectation?
- Given a node v, is the number of links with one end v and load k at a given time well-concentrated around its expectation?
- In a 'transient' situation (when the process starts from a fixed state), are these expectations close to the solution of a differential equation, for a reasonable length of time?
- In equilibrium, are these expectations close to a fixed point of the same differential equation?

We are able to give some answers to these questions.

Our approach is based on couplings and concentration of measure inequalities, and has applications in other settings.

Our work

- We analyse the behaviour of the dynamic model with node set $V_n = \{1, \ldots, n\}$, as defined above, where calls are routed according to any GDAR algorithm. (To be definite, we consider the BDAR algorithm, but our methods extend.)
- $X_t(\{u, v\}, w)$ denotes the number of calls between u and v in progress at time t which are routed via w, i.e. routed along the path consisting of links $\{u, w\}$ and $\{v, w\}$.
- Also, $X_t(\{u, v\}, 0)$ is the number of calls between u and v at time t routed directly.

- $X_t = (X_t(\{u,v\}, w), X_t(\{u,v\}, 0) : \{u,v\} \in E_n, w \in V_n \setminus \{u,v\})$ is the load vector at time t, and takes values in $S = \{0, \ldots, C(n)\}^{N(n-1)}$, where $N = \binom{n}{2}$.
- Given a load vector x and a pair u, v of nodes, let $x(\{u,v\})$ denote the load of link $\{u,v\}$.
- Given a load vector x, node v and $k \in \{0, \ldots, C\}$, let $f_{v,k}(x)$ be the number of links $\{v,w\}$ $(w \neq v)$ in x such that $x(\{v,w\}) = k$.
- Assume we use the BDAR algorithm with d choices, where $d \geq 1$ is fixed or may depend on n.

Differential equation

For a vector $\xi = (\xi(k) : k = 0, \ldots, C)$, let $\xi(\leq j) = \sum_{k=0}^{j} \xi(k)$.

(Think of $\xi(k)$ as the proportion of links with k calls.)

For $0 < k < C$, let

$$\begin{aligned} F_k(\xi) &= \lambda\xi(k-1) - \lambda\xi(k) \\ &\quad + \lambda g_{k-1}(\xi) - \lambda g_k(\xi) \\ &\quad - k\xi(k) + (k+1)\xi(k+1), \end{aligned}$$

where functions g_j are given below.

$$
\begin{aligned}
g_j(\xi) \quad = \quad & 2\xi(C)\xi(j)\xi(\leq j)\sum_{r=1}^{d}(1-\xi(\leq j)^2)^{r-1} \\
& \times (1-\xi(\leq j-1)^2)^{d-r} \\
& + 2\xi(C)\xi(j)\sum_{i=j+1}^{C}\xi(i)\sum_{r=1}^{d}(1-\xi(\leq i)^2)^{r-1} \\
& \times (1-\xi(\leq i-1)^2)^{d-r}.
\end{aligned}
$$

$d = 1$

$$g_j(\xi) \quad = \quad 2\xi(C)(1-\xi(C))\xi(j)$$

Let also

$$\begin{aligned} F_0(\xi) &= -\lambda\xi(0) - \lambda g_0(\xi) + \xi(1); \\ F_C(\xi) &= \lambda\xi(C-1) + \lambda g_{C-1}(\xi) - C\xi(C). \end{aligned}$$

For any initial state $\xi_0 \geq 0$ such that $\sum_k \xi_0(k) = 1$, the differential equation

$$\frac{d\xi_t(k)}{dt} = F_k(\xi_t), \qquad k = 0, 1, \ldots, C$$

has a unique solution, and $\sum_k \xi_t(k) = 1$ for all t.

Given a pair of nodes u, v and an integer $j \in \{0, \ldots, C\}$, let $\mathbb{I}^j_{uv} : S \to \{0, 1\}$ be defined by $\mathbb{I}^j_{uv}(x) = 1$ if $x(\{u, v\}) = j$ and $\mathbb{I}^j_{uv}(x) = 0$ otherwise.

Thus $\mathbb{I}^j_{uv}$ is the indicator of the set of load vectors x where the load of link $\{u, v\}$ is j.

Note that $\mathbb{I}^j_{uv} = \mathbb{I}^j_{vu}$ and that we adopt the convention that $\mathbb{I}^j_{vv}$ is identically 0 for each v and j.

Let functions $\varphi_1, \varphi_2, \varphi_3 : S \to \mathbb{R}$ be defined by

$$\begin{aligned}\varphi_1(x) &= \max_{u,v:u\neq v} \max_{j,k} \Big| \frac{1}{n-2} \sum_{w\in V_n} \mathbb{I}^j_{vw}(x)\mathbb{I}^k_{uw}(x) \\ &\quad - \frac{1}{(n-2)^2} \sum_{w\neq u,v} \mathbb{I}^j_{vw}(x) \sum_{w'\neq u,v} \mathbb{I}^k_{uw'}(x) \Big|;\end{aligned}$$

$$\begin{aligned}\varphi_2(x) &= \max_{u,v:u\neq v} \max_{j} \frac{1}{n-1} |f_{u,j}(x) - f_{v,j}(x)| \\ &= \max_{u,v:u\neq v} \max_{j} \frac{1}{n-1} |\sum_{w\neq u} \mathbb{I}^j_{uw} - \sum_{w\neq v} \mathbb{I}^j_{vw}|.\end{aligned}$$

$$\varphi_3(x) = \max_{u,v:u\neq v} \frac{1}{n-2} \sum_{w\neq u,v} x(\{u,v\},w);$$

Let $\varphi = \max\{\varphi_1, \varphi_2, \varphi_3\}$.

Interpretation of the φ-functions

Functions $\varphi_1, \varphi_2, \varphi_3$ measure how 'uniform' the state of the process is.

- If φ_1 is small, the loads on links around any two nodes are nearly independent.
- If φ_2 is small, any two nodes have similar distributions of loads on the links incident to them.
- If φ_3 is small, there are few indirectly routed calls between any pair of vertices, so the indirectly routed calls are distributed reasonably evenly throughout the network.

Main non-equilibrium result (Luczak 2013+)

For constants λ, d and t_0, there exist constants c_1, c_2 and c_3 such that the following holds.
Let $\xi_0 \in \mathbb{R}^{C+1}$ satisfy $\xi_0(j) \geq 0$ for all j and $\sum_{j=0}^{C} \xi_0(j) = 1$, and let (ξ_t) be the unique solution to the differential equation

$$\frac{d\xi_t}{dt} = F(\xi_t)$$

on $[0, t_0]$, subject to initial condition ξ_0.

Assume that initially there are at most $2\lambda\binom{n}{2}$ calls in the system. Let A be the event that, for each $t \leq t_0$, each $j \in \{0, \ldots, C\}$, and each $v \in V_n$,

$$\begin{aligned} |f_{v,j}(X_t) - n\xi_t(j)| \quad &\leq \quad c_1 \lambda d^2 C^3 t_0 \Big(\sqrt{n} \log n + n\varphi(X_0) \\ &\qquad + \max_{u,j} |f_{u,j}(X_0) - n\xi_0(j)| \Big) e^{c_2 \lambda d^2 C^3 t_0}. \end{aligned}$$

Then, for n large enough, $\mathbb{P}(\overline{A}) \leq e^{-\log^2 n / c_3 C}$.

For the special case $d = 1$ we obtain sharper bounds, replacing the term C^3 in the exponent with C.

Suppose that, for each n, $X_0 = x_0$ a.s. for some deterministic load vector x_0 such that, for some constant c, $\varphi(x_0) \leq \frac{c \log n}{\sqrt{n}}$ and $\max_{v,j} |(n-1)^{-1} f_{v,j}(x_0) - \xi_0(j)| \leq \frac{c \log n}{\sqrt{n}}$. Suppose also that, as $n \to \infty$, λ and t_0 are bounded away from 0, and that $\lambda d^2 C^3 t_0 = o(\log n)$ and $d\lambda t_0 = o(\log \log n)$. Then the theorem implies that, for $\epsilon > 0$, if A^ϵ is the event that, for each $v \in V_n$, each $k \in \{0, \ldots, C\}$, and each $t \in [0, t_0]$,

$$|f_{v,k}(X_t) - (n-1)\xi_t(k)| \leq n^{1/2+\epsilon},$$

then $\mathbb{P}(\overline{A^\epsilon}) \to 0$ as $n \to \infty$.

In the case of λ and C constant, and $d = 1$, our theorem is a more refined, quantitative, version of the law of large numbers in Crametz and Hunt.

Also, our result in this case is related to those in Graham and Meléard, but we do not need to assume that initially all the nodes are exactly exchangeable. Instead, our law of large numbers result holds for a large class of deterministic initial states, and holds simultaneously for all nodes.

The remaining cases of our theorem are completely new.

Our theorem holds also in the case where there is capacity division and direct links are not used (i.e., each arriving call is allocated to the best among d indirect routes), with a suitably modified function F in the differential equation.
Indeed, for $0 < k < C$, we take instead

$$F_k(\xi) = \lambda g_{k-1}(\xi) - \lambda g_k(\xi) - k\xi(k) + (k+1)\xi(k+1),$$

where the functions $g_k(\xi)$ are amended by dropping the factor $\xi(C)$; $F_0(\xi)$ and $F_C(\xi)$ are modified in the same way.
This is important as the model with capacity division does exhibit the power of two choices phenomenon, while the model without capacity division does not.

Initial conditions

What initial conditions allow us to have $\varphi(X_0) \leq cn^{-1/2} \log n$ for some constant $c > 0$, for n large enough?

For example, $X_0 = 0$ works. This is also true, whp, for a state obtained by throwing $\lfloor c\binom{n}{2} \rfloor$ calls into the network at time 0 using the BDAR algorithm, though some work is required to prove this.

Remarks

- The law of large numbers for the BDAR algorithm proved here is valid for the model without direct links in the parameter range considered in L. and McDiarmid (2013+), i.e., with constant λ and d, and $C = C(n) = O(\log \log n)$ for $d \geq 2$, and $C = C(n) = O(\log n / \log \log n)$ for $d = 1$.
- Our methods apply to any GDAR algorithm, and indeed any of the variants discussed above. Obviously, the exact form of the function F in the limiting differential equation will be different for different versions of the model.

Long-term behaviour

We have seen that the process follows the differential equation over a bounded time interval. In some cases, we can show more. Even in the case $d = 1$, the differential equation

$$\frac{d\xi_t}{dt} = F(\xi_t),$$

may have more than one fixed point, i.e., more than one solution to $F(\xi) = 0$. This was observed by Gibbens, Hunt and Kelly: for large enough C, there is a range of λ where there are two stable fixed points. In such a situation, we would not expect to see rapid mixing, or strong concentration of measure in equilibrium.

Rapid mixing

However, if the arrival rate is either sufficiently small or sufficiently large, then the equation has a unique fixed point, and we might expect that the equilibrium of the process is strongly concentrated around this fixed point.

Brightwell and L. consider the cases where $\lambda \leq m_1/d$, and where $\lambda \geq m_2 C^2 d \log(C^2 d))$, for suitable constants m_1, m_2. We prove that, in either of these two regimes, the corresponding sequence of Markov chains is rapidly mixing, and that, for each node v and each $j \in \{0, \ldots, C\}$, $f_{v,j}$ is well concentrated around the fixed point. This establishes a strong form of the 'Erlang fixed point approximation' proposed by Gibbens, Hunt and Kelly, in these regimes.

Main result for equilibrium (Brightwell and Luczak 2013+)

There are constants m_1, m_2 such that, if either $\lambda < m_1/d$, or $\lambda \geq m_2 C^2 d \log(C^2 d)$, then the following hold for sufficiently large n. Here, π denotes the equilibrium distribution of the chain.

- The Markov chain X is rapidly mixing, in time $O(\log n)$.
- There are constants $c_1, c_2 > 0$, depending on d, C and λ, such that, for each node v, each $j \in \{0, \ldots, C\}$, each t, and any $a > 0$,

$$\mathbb{P}_\pi \left(|f_{v,j}(X_t) - \mathbb{E}_\pi f_{v,j}(X_t)| > 2a \right) \leq 3 \exp\left(- \frac{a^2}{c_1 n + c_2 a} \right).$$

- There is a unique solution η^* to the equation $F(\eta) = 0$.
- For all nodes v, all $j \in \{0, \ldots, C\}$, and all t,

$$\left| \frac{1}{n-1} \mathbb{E}_\pi f_{v,j}(X_t) - \eta^*(j) \right| \leq 160 d^2 (C+1)^4 \frac{\log n}{\sqrt{n}}.$$

- Let A be the event that $|f_{v,j}(X_t) - (n-1)\eta^*(j)| \leq 200 d^2 (C+1)^4 \sqrt{n} \log n$, for all nodes v and all $j \in \{0, \ldots, C\}$. Then $\mathbb{P}_\pi(\overline{A}) \leq 3Cn^2 e^{-\delta \log^2 n}$ for some constant $\delta = \delta(d, C, \lambda)$.

Comments

- Our results hold for $\lambda \leq 1/(8d+4)$, when there are rather few calls in the system in equilibrium, so the vast majority of the arriving calls are routed directly.
- In our other regime, with $\lambda \geq m_2 C^2 d \log(C^2 d)$, most links are fully loaded in equilibrium, so most arriving calls are rejected, and extremely few calls are routed indirectly.
- We hope to be able to improve both bounds in future; ideally it should be possible to prove similar results whenever the approximating differential equation has a unique fixed point, but we are far away from that at the moment.

- Our results certainly extend to some other versions of the routing model. One remaining challenge is to cover cases where λ is fixed and C tends to infinity.
- We would also like to prove results about the "bistable" case.

Transient behaviour – concentration of measure

- We derive new concentration of measure inequalities for discrete-time Markov chains based on a version of the bounded differences inequality.
- We analyse a simple coupling of copies of the process to enable us to apply these inequalities. We show that the distance between two coupled copies of the process does not increase too much over a fixed time interval.
- Hence we establish concentration of measure over a fixed time interval for various 'nice' functions of the process, including functions $f_{v,k}(x)$ among others.

Transient behaviour – coupling

The coupling we analyse is as follows.

- For arrivals, calls arrive in both copies of the chain at the same times: calls choose the same endpoints, and the same set of intermediate nodes to consider.
- For departures, calls are paired up as far as possible, so that a call present in both copies of the chain departs at the same time.
- Departures never increase the distance between the copies, and sometimes decrease it; arrivals may increase the distance.

- Let A be the generator operator of the Markov process X. By standard theory of Markov chains, for each $t \geq 0$,

$$\frac{d\,\mathbb{E}[f_{v,k}(X_t)]}{dt} = \mathbb{E}[Af_{v,k}(X_t)].$$

- We use our concentration of measure estimates, as well as the fact that all the nodes are exchangeable, to approximate the expected drift of the functions $f_{v,k}(X_t)$, i.e., the $\mathbb{E}[Af_{v,k}(X_t)]$, in terms of polynomial functions of $\mathbb{E}[f_{v,j}(X_t)]$ for the various j.

Let ζ_t^v be the vector with components $\zeta_t(v,j) = (n-1)^{-1}\,\mathbb{E}[f_{v,j}(X_t)]$. Let $\widetilde{\varphi} = \max(\varphi_1, \varphi_2)$.

We have, for all $v \in V_n$ and $j \in \{0, \ldots, C\}$,

$$\left|\mathbb{E}[g_{v,j}(X_t)] - (n-1)g_j(\zeta_t^v)\right| \leq 12d^2(C+1)^3 n\,\mathbb{E}[\widetilde{\varphi}(X_t)] + 20d^2(C+1)\sqrt{n}\log n.$$

All the other terms in the expression for $Af_{v,k}(X_t)$ are linear.

Hence the above inequality provides us with a bound on $|\,\mathbb{E}[Af_{v,k}(X_t)] - F_k(\zeta_t)|$, in terms of $\mathbb{E}[\widetilde{\varphi}(X_t)]$.

- We thus obtain an approximate differential equation satisfied by $(n-1)^{-1}\,\mathbb{E}[f_{v,k}(X_t)]$.
- We use concentration of measure estimates again to argue that $(n-1)^{-1} f_{v,k}(X_t)$ stays close to the k-th component of the solution uniformly over an interval.

Equilibrium behaviour

- For sufficiently small arrival rate λ, we show that the coupling used above is actually contractive: the distance decreases in expectation over time. The same is true if λ is sufficiently large, provided the chain remains in a "good set" of states.
- We then show that, in either of these two regimes, the chain exhibits rapid mixing.

- We use our concentration of measure inequalities, as well as the fact that the coupling is contractive, to establish strong concentration of measure, with uniform bounds over all time, for the process starting from a fixed state.
- We deduce strong concentration of measure for "nice" functions of the process X_t in equilibrium.

- Recall that π denotes the equilibrium distribution of our chain.
- In equilibrium, we have

$$0 = \frac{d\,\mathbb{E}_\pi[f_{v,k}(X_t)]}{dt} = \mathbb{E}_\pi[Af_{v,k}(X_t)].$$

- As in the transient case, we bound $|\,\mathbb{E}_\pi[Af_{v,k}(X_t)] - F_k(\zeta)|$ in terms of $\widetilde{\varphi}$, where $\zeta(j) = \frac{1}{n-1}\,\mathbb{E}_\pi\, f_{v,j}(X_t)$.
- We show that, in equilibrium, the expectations of $\varphi_1(X_t)$ and $\varphi_2(X_t)$ are small.

- We show that, in either of the two regimes we consider, with the arrival rate either sufficiently small or sufficiently large, the equation $F(\xi) = 0$ has a unique solution, and that any "approximate solution" to the equation lies close to the actual solution.
- We deduce that, in either of our two regimes, each $f_{v,k}(X_t)$ is strongly concentrated around the kth component of the unique solution to $F(\xi) = 0$.

Concentration of measure inequalities

In order to analyse this process, it was necessary to develop new concentration of measure inequalities. We shall state and discuss two such inequalities. The first sets the scene, but, as we shall discuss, it does not suit our intended application.

First inequality

Theorem (L., 2013+)

Let P be the transition matrix of a discrete-time Markov chain with discrete state space S. Let $f : S \to \mathbb{R}$ be a function. Let $(\alpha_i : i \in \mathbb{Z}^+)$ be a sequence of positive constants such that for all $i \in \mathbb{Z}$,

$$\sup_{x,y \in S : P(x,y) > 0} |\mathbb{E}_{\delta_x} P^i(f) - \mathbb{E}_{\delta_y} P^i(f)| \leq \alpha_i.$$

Then for all $u > 0$, $x_0 \in S$, and $t > 0$,

$$\mathbb{P}_{\delta_{x_0}}(|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u) \leq 2e^{-u^2/2(\sum_{i=0}^{t-1} \alpha_i^2)}.$$

More generally, let S_0 be a non-empty subset of S, and let $(\alpha_i : i \in \mathbb{Z})$ be a sequence of positive constants such that, for all $i \in \mathbb{Z}$,

$$\sup_{x,y \in S_0 : P(x,y) > 0} |\, \mathbb{E}_{\delta_x P^i}(f) - \mathbb{E}_{\delta_y P^i}(f)| \leq \alpha_i.$$

Let

$$S_0^0 = \{x \in S_0 : y \in S_0 \text{ whenever } P(x,y) > 0\}.$$

Then for all $x_0 \in S_0^0$, $u > 0$ and $t > 0$,

$$\mathbb{P}_{\delta_{x_0}} \left(\{|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u\} \cap \{X_s \in S_0^0 : \ 0 \leq s \leq t\} \right) \leq 2e^{-u^2/2(\sum_{i=0}^{t-1} \alpha_i^2)}.$$

- In a typical application that I have in mind, there would be a sequence of such Markov chains, indexed by n.
- We would be interested in functions with expectation of order about n, and would run the chain for about n steps.
- Also, we could take $\alpha_i = \alpha_i^{(n)} = (\alpha^{(n)})^i$, where $\alpha^{(n)} \leq 1 + c/n$ and $c > 0$.

- We would then obtain a concentration of measure inequality of the form

$$\mathbb{P}_{\delta_{x_0}}(|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u) \leq 2e^{-u^2/c_1 n}, \quad t \leq c_2 n.$$

- Unfortunately, for our model, we would be in trouble: we want to run the process for $c_2 n^2$ steps, and we want concentration for functions of order n.
- We have a more sophisticated version of the inequality, useful when a good uniform bound on $\sup_{x,y \in S: P(x,y)>0} |\mathbb{E}_{\delta_x} P^i(f) - \mathbb{E}_{\delta_y} P^i(f)|$ fails to hold.

Second inequality

Theorem

Let $f : S \to \mathbb{R}$ be a function. Suppose the set S_0 and numbers $\alpha_{x,i}(y)$ $(x, y \in S_0)$ are such that, for all $i \in \mathbb{Z}^+$ and all $x, y \in S_0$,

$$|\,\mathbb{E}_{\delta_x P^i}(f) - \mathbb{E}_{\delta_y P^i}(f)| \leq \alpha_{x,i}(y).$$

Let

$$S_0^0 = \{x \in S_0 : y \in S_0 \text{ whenever } P(x, y) > 0\}.$$

Assume that, for some sequence $(\alpha_i : i \in \mathbb{Z}^+)$ of positive constants,

$$\sup_{x \in S_0^0} (Pa_{x,i}^2)(x) \leq \alpha_i^2.$$

Let $t > 0$, and let $\beta = 2\sum_{i=0}^{t-1} \alpha_i^2$. Suppose also that $\hat{\alpha}$ is such that

$$\sup_{0 \leq i \leq t-1} \sup_{x,y \in S_0^0, P(x,y)>0} \alpha_{x,i}(y) \leq \hat{\alpha}.$$

Finally, let $A_t = \{\omega : X_s(\omega) \in S_0^0 : 0 \leq s \leq t\}$.
Then, for all $u > 0$,

$$\mathbb{P}_{\delta_{x_0}} \left(\left\{ |f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u \right\} \cap A_t \right) \leq 2e^{-u^2/(2\beta(1+(2\hat{\alpha}u/3\beta))}.$$

We next state a result from McDiarmid (1998), illustrating the "bounded differences" approach.
Our first concentration inequality can be derived from this result in a relatively straightforward way. Our second concentration inequality is derived from a different result in the same article. The inequalities we state above are particularly suited to applications, as we hope to illustrate here.

Let $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ be a probability space, with $\widetilde{\Omega}$ finite. Let $\widetilde{\mathcal{G}} \subseteq \widetilde{\mathcal{F}}$ be a σ-field of subsets of $\widetilde{\Omega}$. Then there exist disjoint sets $\widetilde{G}_1, \ldots, \widetilde{G}_m$ such that $\widetilde{\Omega} = \cup_{r=1}^{m} \widetilde{G}_r$ and every set in $\widetilde{\mathcal{G}}$ can be written as a union of some of the sets $\widetilde{G}_r$.
Given a bounded random variable Z on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$, the *conditional supremum* $\sup(Z \mid \widetilde{\mathcal{G}})$ of Z in $\widetilde{\mathcal{G}}$ is given by

$$\sup(Z \mid \widetilde{\mathcal{G}})(\widetilde{\omega}) = \min_{\widetilde{A} \in \widetilde{\mathcal{G}} : \widetilde{\omega} \in \widetilde{A}} \max_{\widetilde{\omega}' \in \widetilde{A}} Z(\widetilde{\omega}') = \max_{\widetilde{\omega}' \in \widetilde{G}_r} Z(\widetilde{\omega}'),$$

where $\widetilde{\omega} \in \widetilde{G}_r$. Thus $\sup(Z \mid \widetilde{\mathcal{G}})$ takes the value at $\widetilde{\omega}$ equal to the maximum value of Z over the event $\widetilde{G}_r$ in $\widetilde{\mathcal{G}}$ containing $\widetilde{\omega}$.

The *conditional range* $\mathrm{ran}(Z)$ of Z in $\widetilde{\mathcal{G}}$ is the $\widetilde{\mathcal{G}}$-measurable function $\mathrm{ran}(Z \mid \widetilde{\mathcal{G}}) = \sup(Z \mid \widetilde{\mathcal{G}}) + \sup(-Z \mid \widetilde{\mathcal{G}})$, that is, for $\widetilde{\omega} \in \widetilde{G}_r$,

$$\mathrm{ran}(Z \mid \widetilde{\mathcal{G}})(\widetilde{\omega}) = \max_{\widetilde{\omega}_1, \widetilde{\omega}_2 \in \widetilde{G}_r} |Z(\widetilde{\omega}_1) - Z(\widetilde{\omega}_2)|.$$

Let $t \in \mathbb{N}$, let $\{\emptyset, \widetilde{\Omega}\} = \widetilde{\mathcal{F}}_0 \subseteq \widetilde{\mathcal{F}}_1 \subseteq \ldots \subseteq \widetilde{\mathcal{F}}_t$ be a filtration in $\widetilde{\mathcal{F}}$, and let $Z_0, \ldots, Z_t$ be the martingale defined by $Z_i = \widetilde{\mathbb{E}}(Z|\widetilde{\mathcal{F}}_i)$ for each $i = 0, \ldots, t$. For each i, let ran_i denote $\mathrm{ran}(Z_i|\widetilde{\mathcal{F}}_{i-1})$; For each j, let R_j^2 be the random variable $\sum_{i=1}^{j} \mathrm{ran}_i^2$, and set

$$\widehat{r}_j^2 = \sup_{\widetilde{\omega} \in \widetilde{\Omega}} R_j^2(\widetilde{\omega}).$$

Theorem (McDiarmid (1998))

Let Z be a bounded random variable on a finite probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ with $\widetilde{\mathbb{E}}(Z) = m$. Let $\{\emptyset, \widetilde{\Omega}\} = \widetilde{\mathcal{F}}_0 \subseteq \widetilde{\mathcal{F}}_1 \subseteq \ldots \subseteq \widetilde{\mathcal{F}}_t$ be a filtration in $\widetilde{\mathcal{F}}$, and assume that Z is $\widetilde{\mathcal{F}}_t$-measurable. Then for any $a \geq 0$,

$$\widetilde{\mathbb{P}}(|Z - m| \geq a) \leq 2e^{-2a^2/\widehat{r}_t^2}.$$

A natural way to verify the hypotheses of our concentration inequalities is to construct couplings between copies of the Markov chain starting from different states, and show that they grow closer together in expectation, with regard to a suitable notion of distance.

(In the transient regime, we only have to show that the chains do not grow too much further apart over a fixed time interval.)

We illustrate this in our application, in the case where the arrival rate λ is small.

We work with a discretised version $(\widehat{X}_t)$ of our continuous time chain. In this chain, the next event is an arrival with probability $\frac{\lambda}{\lambda+C}$, and a "potential departure" with probability $\frac{C}{\lambda+C}$.

Conditioned on the event being an arrival, the arriving call is routed as in the BDAR algorithm.

The calls in progress are numbered with distinct numbers from $\{1, \ldots, C\binom{n}{2}\}$. Conditioned on the event being a departure, a uniform random number from this set is chosen, and if there is a call with that number, it departs.

Given two copies $(\widehat{X}_t)$ and $(\widehat{Y}_t)$ of our chain, we couple them in a natural way. Departures and arrivals coincide for the two chains.

If the event is to be an arrival, the same endpoints are chosen in both chains, and the same list of intermediate nodes.

If the event is to be a departure, then calls are numbered so that, as far as possible, calls on the same route in both chains are given the same number, so that they depart together.

We assume that our two coupled copies $(\widehat{X}_t)$ and $(\widehat{Y}_t)$ differ by one call at time t, present in $\widehat{X}_t$ but not in $\widehat{Y}_t$.

We first consider the natural ℓ_1-distance between two states,

$$\begin{aligned}\|x-y\|_1 &= \sum_{u,v} |x(\{u,v\},0) - y(\{u,v\},0)| \\ &\quad + \sum_{\{u,v\},w} |x(\{u,v\},w) - y(\{u,v\},w)|\end{aligned}$$

We look at the change in expected distance between the two copies after one step of the discrete chain.

Departures are straightforward to analyse: all calls are paired except the one extra call in $\widehat{X}_t$. The departure of this extra call decreases the distance from 1 to 0. The departure of any other pair of calls, one in $\widehat{X}_t$ and one in $\widehat{Y}_t$, does not change the distance.

Most of the time, an arrival does not change the distance. However, if one of the links whose load is inspected carries the extra call, then the arriving call may be routed differently in the two chains, so the distance could go up, by at most 2.

The conditional probability of such a "bad arrival" is at most $\frac{2\lambda(2d+1)/\binom{n}{2}}{\lambda+C}$, since only $2d+1$ links are inspected for possible routing of the arrival call, and there are at most 2 links carrying the extra call.

So the change in the expected distance between the chains on one step of the coupled processes is at most

$$\frac{1}{\binom{n}{2}(\lambda + C)}\big(-1 + 4\lambda(2d+1)\big),$$

which is negative if $\lambda < 1/(8d+4)$.

This is enough to establish that, in this regime, the chain is rapidly mixing, in $O(n^2 \log n)$ steps of the discrete chain, corresponding to time $O(\log n)$ in the continuous chain.

But we need more to apply our concentration inequality.

For each node v, we consider the "local distance"

$$\begin{aligned}\|x-y\|_v &= \sum_u |x(\{u,v\},0) - y(\{u,v\},0)| \\ &\quad + \sum_{u,w} |x(\{u,w\},v) - y(\{u,w\},v)| \\ &\quad + \sum_{u,w} |x(\{v,w\},u) - y(\{v,w\},u)|.\end{aligned}$$

We show that

$$\begin{aligned}\mathbb{E}(\|\widehat{X}_{t+1} - \widehat{Y}_{t+1}\|_v \mid \widehat{X}_t, \widehat{Y}_t) &\leq \left(1 - \frac{1-(8d+4)\lambda}{(\lambda+C)\binom{n}{2}}\right)\|\widehat{X}_t - \widehat{Y}_t\|_v \\ &\quad + \frac{\lambda}{\lambda+C}\frac{12d^2}{\binom{n}{2}(n-2)}\|\widehat{X}_t - \widehat{Y}_t\|_1.\end{aligned}$$

A simple induction argument leads to

$$\mathbb{E}(\|\widehat{X}_t - \widehat{Y}_t\|_v \mid \widehat{X}_0, \widehat{Y}_0)$$
$$\leq \left(1 - \frac{1-(8d+4)\lambda}{(\lambda+C)\binom{n}{2}}\right)^t \left(\|\widehat{X}_0 - \widehat{Y}_0\|_v + \frac{50d^2\lambda t}{(\lambda+C)n^3}\|\widehat{X}_0 - \widehat{Y}_0\|_1\right).$$

Now, for a function f satisfying $|f(x) - f(y)| \leq \|x - y\|_v$, we can take

$$\alpha_{x,i}(y) = \left(1 - \frac{1 - (8d+4)\lambda}{(\lambda + C)\binom{n}{2}}\right)^i \left(\|x - y\|_v + \frac{50d^2\lambda i}{(\lambda + C)n^3}\|x - y\|_1\right).$$

The key is that, for any x and any $i \geq 0$, if y is chosen with probability $P(x, y)$, then it is very likely that $\|x - y\|_v = 0$, and thus $\alpha_{x,i}(y)$ is relatively small.
Hence we can provide good bounds on $P\alpha_{x,i}^2$, and use the full power of our second inequality.

High arrival rate

- If the arrival rate λ is high, we first show that the process rapidly enters a set of states where most links are fully loaded.
- We also define a tailored distance function, and show that, under the coupling, the distance between the two states decreases in expectation, provided the processes both stay in the good set.

One fixed point?

As observed by Gibbens, Hunt and Kelly (1990), the differential equation above does not always have a unique fixed point, even in the case $d = 1$. The found a value of C, around 200, and a small range of λ, with λ of similar magnitude to C, where there are two attractive fixed points of the differential equation.

Our results are very unlikely to hold as they stand in a regime where the differential equation has more than one attractive fixed point.

The ranges we treat, where λ is either sufficiently small or sufficiently large, are far away from the range discovered by Gibbens, Hunt and Kelly. In our range, it is quite easy to prove that there is a unique fixed point.