# Globs in the Primordial Soup

## The Emergence of Connected Crowds in Mobile Wireless Networks

Simon Heimlicher
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
heimlicher@tik.ee.ethz.ch

Kavé Salamatian
LISTIC-PolyTech
Université de Savoie, France
kave.salamatian@univ-savoie.fr

Extended Version

- Includes proofs → Sec. 2
- Further results from synthetic random walk scenario → Sec. 3.2

## ABSTRACT

In many practical scenarios, nodes gathering at points of interest yield sizable connected components (clusters), which sometimes comprise the majority of nodes. While recent analysis of mobile networks focused on the process governing node encounters ("contacts"), this model is not particularly suitable for gathering behavior. In this paper, we propose a model of stochastic coalescence (merge) and fragmentation (split) of clusters. We implement this process as a Markov chain and derive analytically the exact stationary distribution of cluster size. Further, we prove that, as the number of nodes grows, the clustering behavior converges to a mean field, which is obtained as a closed-form expression. This expression translates the empirical merge and split rate of a scenario, a microscopic property, to an important macroscopic property—the cluster size distribution—with surprising accuracy. We validate all results with synthetic as well as real-world mobility traces from conference visitors and taxicabs with several thousand nodes.

## 1. INTRODUCTION

Up to now, the analysis of mobile networks was predominantly based on modeling individual nodes or node encounters. In particular, research on delay-tolerant networks (DTN) made progress in characterizing the stochastic process underlying single-hop paths ("contacts") and leveraging emerging "space-time paths" for communication in a disconnected network. Those studies laid an important foundation toward understanding the temporal characteristics of contacts and designing contact-based forwarding schemes for delay-tolerant applications tailor-made for those networks.

Yet, we argue that in addition to space-time paths, multi-hop paths may often exist in mobile wireless networks. Such multi-hop paths may enable some of the most popular applications, such as web browsing and email, in mobile wireless networks, even though the protocols used by those applications (in particular TCP [19]) have not been designed to be delay-tolerant. We argue that understanding the circumstances that lead to multi-hop paths is paramount to design routing and forwarding algorithms that leverage those communication opportunities. In our discussion, we distinguish between *partial paths*, which bridge only part of the way from the source to the destination, and *full paths* connecting source and destination; and we use the term *partially-connected network* to refer to a network that is disconnected but provides partial paths.

Following those definitions, one may ask if disconnected networks with partial paths exist. Continuum percolation theory [12] is often used to study connectivity; its main result is that if the node density is below the percolation threshold, almost all nodes are isolated; above the threshold, the network "percolates" and forms one giant connected component. Yet, [32] shows that in clustered networks, such a phase transition does not occur. In delay-tolerant networking, characterization of the contact process [9, 15] is mostly concerned with forwarding based on individual contacts, though there are proposals to leverage multi-hop paths [28] as well.

In light of this, to analyze rigorously the existence and characteristics of partial paths, a new methodology seems in order. Node independence is at the core of most analytical models for mobile networks, but multi-hop paths are at odds with this assumption. Fortunately, if the behavior of individual nodes is abstracted from, the dependency between nodes on a multi-hop path can be hidden. Under this premise, one could consider multi-hop paths as the basic entities, but those may still be connected to each other. Therefore we propose to lump together all connected multi-hop paths to connected components (clusters) and describe a mobile scenario through merge and split events between such clusters.

In this paper, we introduce a model for arbitrary mobile wireless networks based on the concept of stochastic coagulation and fragmentation [2]. In analogy to globs of particles coalescing and fragmenting in a system of particles in a solvent, we describe a system of $N$ mobile nodes as a set of clusters that merge and split. The state of the system is represented by a vector whose elements $i = 1, 2, \ldots$ are the number of clusters of size $i$. As $N$ is constant, two primitive events may occur in this system. First, two clusters of sizes $k$ and $l$ may merge into a new cluster of size $k + l$, as described by the merge process. Second, a cluster of size $k + l$ may split into two clusters of sizes $k$ and $l$, according to the split process. It follows that the merge and the split process determine the stationary distribution of cluster size and thus, whether the network is connected, disconnected, or partially connected by partial paths.

We implement this model as a Markov process over the finite state space of all partitions of $N$. Under certain conditions, this process is reversible and thus its stationary distribution (corresponding to the distribution of cluster size) is obtained in closed form. Furthermore, we prove that the behavior of the merge–split process converges to a mean field for large numbers of nodes, $N$. The mean field is obtained in closed form and provides a remarkably precise approximation even for realistic networks with finite numbers of nodes. Indeed, with increasing numbers of nodes, numerical errors undermine computation of the exact distribution and the mean field approximation may be preferable for large networks.

Since the versatility of the merge–split model may not be apparent, we also aim to illustrate applications to the extent space permits. The mean field approximation intuitively translates through a simple expression a microscopic property——the three parameters of the merge-split process——to the macroscopic cluster size distribution. Thus, our analysis characterizes mobile network scenarios uniquely by these three parameters and provides an accessible interpretation through the cluster size distribution. In particular, this also indicates to what extent a scenario provides partial paths, which gives important insights into the structure of the network and in particular its suitability for contact-based [21], hop-by-hop, or end-to-end forwarding [17] schemes.

The prediction quality of the mean field approximation as well as the exact distribution are validated against random walk simulation as well as mobility traces. We use traces from conference visitors and taxicabs in San Francisco and Shanghai. More specifically, we extract the merge and the split rate from these scenarios and then derive the cluster size distribution using both our exact analytical solution and the mean field approximation. The exact result yields a remarkably precise prediction; the mean field approximation by its nature gives a reliable prediction of the shape of the distribution; in particular it predicts whether giant components emerge.

The characterization of mobile networks via the stationary distribution we present in this paper is the first of several applications of the merge–split process; we outline further results at the end of this paper.

The contributions we present in this paper are as follows.

- We introduce a Markov process modeling merge and split behavior of connected components (clusters) in mobile wireless networks;
- we derive analytically its stationary distribution, corresponding to the distribution of cluster size;
- we prove the convergence of the stationary distribution to a mean field behavior with increasing number of nodes,
- the mean field approximation is a closed-form expression that translates the microscopic behavior (merge–split process) to an important macroscopic property, the cluster size distribution;
- we validate the quality of the prediction of cluster size distribution (exact derivation and mean field approximation) against random walk mobility and three real-world traces.

In the next section, we introduce the merge–split model and prove its convergence to a mean field behavior; we further outline the calibration of the model with empirical data. We validate the exact derivation as well as the mean field approximation in Sec. 3 by comparing the predicted against the empirical cluster size distribution in synthetic and real-world mobility traces. In Sec. 4, we discuss our contributions vis-a-vis related work. Finally, in Sec. 5 we conclude the paper, giving a preview of results that are to be published, as well as other future work.

# 2. ANALYTIC FORMULATION

In this section we describe the analytical model we use throughout. We then analyze its convergence to a mean field behavior in Sec. 2.2. Finally, Sec. 2.3 describes how the model can be calibrated to match an empirical dataset and discusses the effect of its parameters.

## 2.1 Finite size system formulation

We describe an arbitrary mobile network as a system of $N$ interacting nodes. At every time $t$, a node is in exactly one cluster, *i.e.*, it is member of a set of nodes that are connected by a full path at time $t$. The state of the system is described by the cluster size vector $(\nu_N(1,t), \nu_N(2,t), \ldots, \nu_N(N,t))$ with elements $\nu_N(i,t)$ representing the number of clusters of size $i$ at time $t$. We consider two primitive interactions between these nodes.

1. **Merge reaction:** A cluster of $k$ nodes merges with a cluster of $l$ nodes, yielding a cluster of $k+l$ nodes:
$$C_k + C_l \to C_{k+l}.$$
This reaction is also called *coalescence* and happens at a rate $\psi_N(k,l)$, which is assumed to be symmetric, *i.e.*, $\psi_N(k,l) = \psi_N(l,k)$. A merge reaction of clusters of sizes $k$ and $l$ has the following drift effect on the cluster size vector:
$(\ldots, \nu_N(k,t)-1, \ldots, \nu_N(l,t)-1, \ldots, \nu_N(k+l,t)+1, \ldots)$.
2. **Split reaction:** A cluster of size $l$ splits into two clusters of sizes $k$, $(k < l)$ and $l-k$:
$$C_l \to C_k + C_{l-k}.$$
This reaction is also called *fragmentation* and happens at a rate $\phi_N(l|k)$ and we assume $\phi_N(l|k) = \phi_N(l|l-k)$. A split reaction of a cluster of size $l$ into two clusters of sizes $l-k$ and $k$ has the below drift effect on the cluster size vector:
$(\ldots, \nu_N(l-k,t)+1, \ldots, \nu_N(k,t)+1, \ldots, \nu_N(l,t)-1, \ldots)$.

We call such a process a merge–split process. These reactions happens subject to the node conservation condition:

$$\sum_{k=1}^{N} k\nu_N(k,t) = N, \forall t \leq 0. \tag{1}$$

This defines a Markov process over the finite state space $\Omega = \Omega_N = \{\tau\}$ of all partitions of $N$. A special case with only the merge reaction is called a Marcus-Lushnikov process [26, 25] and has gained attention from the mathematical community. The analogous process with only the split reaction is called a fragmentation process and has been studied extensively in the context of branching processes. The problem we analyze here is a mix of these two problems.

To simplify the notation, we will drop index $N$ referring to the total number of nodes in the forthcoming unless needed. As we will see it is useful to define the following intensity ratio function $q(k,l)$, based on the ratio between merge and split intensity, as:

$$q(k,l) = \begin{cases} \dfrac{\psi(k,l)}{\phi(k+l|l)}, & \text{if } \psi(k,l)\phi(k+l|l) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The existence of a stationary equilibrium state is conditioned on the reversibility of the Markov chain; a Markov process $\mathcal{M}_t$ is said to be reversible with respect to a probability measure $\mu$ if for all

$t \geq 0$, the processes $\mathcal{M}_t^\mu$, $0 \leq s \leq t$ and $\mathcal{M}_{t-s}^\mu$, $0 \leq s \leq t$, starting from the same initial distribution $\mu$, have the same finite dimensional distribution [22]. Reversibility is an important property of a Markov process; if a reversible process is ergodic, its unique stationary distribution is the reversible measure. The reversible measure can be derived in general using the flow equilibrium equation of the Markov chain, *i.e.*, $V(\tau,\xi)\mu(\tau) = V(\xi,\tau)\mu(\xi)$, where $V(\tau,\xi)$ is the total intensity of transitions from state $\tau$ to $\xi$.

The following theorem gives necessary and sufficient conditions under which the process is reversible and will therefore converge to a stationary equilibrium state.

THEOREM 1. *[22] Suppose that $q(k,l) > 0$, for $2 < k+l < N$, then the merge–split Markov process is reversible if and only if for some function $a(k) > 0,\ k = 1,\ldots,N$, we can rewrite $q(k,l)$ as*

$$q(k,l) = \frac{a(k+l)}{a(k)a(l)}.$$

This theorem leads to the well-known deterministic detailed balance condition on intensities [22], *i.e.*,

$$a(k)a(l)\psi(k,l) = a(k+l)\phi(k+l|l),\ k,l : 2 \leq k+l \leq N$$

for some $a(k) > 0,\ k = 1,\ldots,N, l = 1,\ldots,N$.

Moreover, a merge–split process will have a reversible distribution $\mu$ following the closed form formula derived as:

THEOREM 2. *Suppose that $q(k,l)$ satisfies the condition given in Thm. 1, then the merge–split process defined above is reversible with respect to the invariant measure $\mu = \mu_N \in \Omega_N$, given by*

$$\mu_N(\tau) = \mathcal{C}_N \frac{a(1)^{n_1} a(2)^{n_2} \ldots a(N)^{n_N}}{n_1! n_2! \ldots n_N!}. \tag{3}$$

*where $\tau(t) = (n_1,\ldots,n_N) \in \Omega_N$ is an acceptable configuration with $n_k$ clusters of size $k$. $\mathcal{C}_N$ is a scaling coefficient defined such that $\sum_{\tau\in\Omega_N} \mu(\tau) = 1$.*

The proof of the theorem proceeds by validating that this distribution satisfies the flow equilibrium condition, which is given by $V(\tau,\xi)\mu(\tau) = V(\xi,\tau)\mu(\xi)$, $\tau,\xi \in \Omega_N$. In the forthcoming we will denote

$$c_N = \frac{1}{\mathcal{C}_N} = \sum_{\tau\in\Omega_N} \frac{a(1)^{n_1} a(2)^{n_2} \ldots a(N)^{n_N}}{n_1! n_2! \ldots n_N!}. \tag{4}$$

The invariant measure gives the stationary state occupation measure, *i.e.*, the probability that the Markov chain is in state $\tau$ in equilibrium. However, for our purpose, we are interested in knowing the statistics of $n_k(\tau)$, *i.e.*, the number of clusters of size $k$ in the configuration $\tau$. The below statistics are of interest:

$$\begin{aligned}
\nu_N(k) &= \mathbb{E}\left\{n_k(\tau)\right\}, && k = 1,\ldots N\\
\varsigma_N(k,l) &= \mathbb{C}\mathrm{ov}\left\{n_k(\tau), n_l(\tau)\right\}, && k \neq l, k, l = 1,\ldots,N\\
\sigma_N^2(k) &= \mathbb{V}\mathrm{ar}\left\{n_k(\tau)\right\}, && k = 1,2,\ldots,N
\end{aligned}$$

The next theorem derives those statistics:

THEOREM 3. *Let $\mu_N$ be given as in Thm. 2, then:*

$$\begin{aligned}
\nu_N(k) &= a(k)\frac{c_{N-k}}{c_N},\\
\varsigma_N(k,l) &= a(k)a(l)\left(\frac{c_{N-k-l}}{c_N} - \frac{c_{N-k}c_{N-l}}{c_N^2}\right),\ k \neq l,\\
\sigma_N^2(k) &= a^2(k)\left(\frac{c_{N-2k}}{c_N} - \frac{c_{N-k}^2}{c_N^2}\right) + a(k)\frac{c_{N-k}}{c_N},
\end{aligned}$$

*for $k,l = 1,\ldots,N$, and $c_{-m} = 0,\ m = 1,\ldots$.*

PROOF. First, denote an acceptable partition of the $N$ nodes into $n_k^{(N)}$ clusters of size $k$ by $\tau^{(N)} = (n_1^{(N)},\ldots,n_N^{(N)})$. Then assume that one puts aside $k$ nodes and has a split–merge problem with $N-k$ nodes. Now, one can split the set of acceptable partitions $\Omega_N$ into two separate subsets $\Omega_N^{n_k\neq 0} = \{\tau \in \Omega_N | n_k > 0\}$ and $\Omega_N^{n_k=0} = \{\tau \in \Omega_N | n_k = 0\}$, such that $\Omega_N = \Omega_N^{n_k\neq 0} \bigcup \Omega_N^{n_k=0}$. One can define a one-to-one equivalence relation between a partition $\tau^{(N-k)} \in \Omega_{N-k}$ and a partition $\tau^{(N)} \in \Omega_N^{n_k\neq 0}$ such that $n_l^{(N)} = n_l^{(N-k)},\ l \neq k$, $n_k^{(N)} = n_k^{(N-k)} + 1$. Using this equivalence relation one can rewrite (4) as:

$$c_N = \sum_{\tau\in\Omega_N^{n_k\neq 0}} \frac{a(1)^{n_1^{(N)}} \cdots a(k)^{n_k^{(N)}} \cdots a(N-k)^{n_{N-k}^{(N)}}}{n_1^{(N)}! \cdots n_k^{(N)}! \ldots n_{N-k}^{(N)}!} +$$
$$\sum_{\tau\in\Omega_N^{n_k=0}} \frac{a(1)^{n_1^{(N)}} \cdots a(k)^0 \ldots a(N)^{n_N^{(N)}}}{n_1^{(N)}! \cdots 0! \cdots n_N^{(N)}!}$$

where the limitation of the first summing term to clusters of size $N-k$ is coming from the existence of at least a cluster of size $k$. Now we have:

$$\frac{\partial c_N}{\partial a(k)} =$$
$$\sum_{\tau\in\Omega_N^{n_k\neq 0}} n_k^{(N)} \frac{a(1)^{n_1^{(N)}} \cdots a(k)^{n_k^{(N)}-1} \cdots a(N-k)^{n_{N-k}^{(N)}}}{n_1^{(N)}! \cdots n_k^{(N)}! \cdots n_{N-k}^{(N)}!}$$

By using the one-to-one equivalence between $\Omega_{N-k}$ and $\Omega_{N-k}^{n_k\neq 0}$ one can rewrite the above equation as:

$$\frac{\partial c_N}{\partial a(k)} =$$
$$\sum_{\tau\in\Omega_{N-k}} \frac{a(1)^{n_1^{(N-k)}} \cdots a(k)^{n_k^{(N-k)}} \cdots a(N-k)^{n_{N-k}^{(N-k)}}}{n_1^{(N-k)}! \cdots n_k^{(N-k)}! \cdots n_{N-k}^{(N-k)}!}$$

showing that $\frac{\partial c_N}{\partial a(k)} = c_{N-k}$.

Moreover, applying the derivative to (4) yields:

$$\frac{\partial c_N}{\partial a(k)} = \frac{c_N}{a(k)} \sum_{\tau\in\Omega_N} \nu_\tau \mu_N(\tau) = \frac{c_N \nu_N(k)}{a(k)}$$

resulting in $\nu_N(k) = a(k)\frac{c_N - k}{c_N}$.

The expression for $\varsigma_N(k,l)$ results from differentiating the above expression with respect to $a(l)$, $l \neq k$. The expression for $\sigma_N^2(k)$ is obtained by differentiating one more time the above expression by $a(k)$. □

Theorem 3 gives a characterization of the distribution of cluster sizes. The correlation between cluster sizes is a result of the finite value of $N$ and the constraint given in (1). In order to complete the characterization, we need to obtain the values $\{c_n\}$.

These values can be derived using the series

$$S(x) = \sum_{i=1}^{\infty} a(i) x^i,$$

which is assumed to converge for $x \in D_S = \{x \mid |x| < R_S\}$ where $R_S$ is the convergence radius.

THEOREM 4. *Under the assumption of convergence of the series $S(x)$ in $D_S$*

1. *The values $c_n$, $n = 1, 2, \ldots$ are the coefficients of the Taylor expansion of the function $g(x) = e^{S(x)}$,* i.e.*,*

$$g(x) = e^{S(x)} = \sum_{n=0}^{\infty} c_n x^n$$

*where $g(x)$ converges over $D_g = D_S$.*

2. *The radii of convergence of the Taylor series of $g(x)$ and $S(x)$ are equal,* i.e.*,*

$$\lim_{n\to\infty} \frac{c_n}{c_{n+1}} = \lim_{n\to\infty} \frac{a_n}{a_{n+1}} = R.$$

3. *The values $c_n$ can be derived by the recurrence relation: $c_0 = 1$, $c_1 = a(1)$,*

$$(n+1)c_{n+1} = \sum_{k=0}^{n} (k+1)a(k+1)c_{n-k},\ n = 1, 2, \ldots$$

PROOF. We can write $e^{S(x)}$ as

$$e^{S(x)} = e^{a(1)x} e^{a(2)x^2} \cdots e^{a(k)x^k} \cdots, \quad x \in D_g.$$

By expanding every term of the right-hand side as a Taylor series, *i.e.*, $e^{a(k)x^k} = \sum_{j=1}^{\infty} \frac{a(k)^j x^{kj}}{(j!)}$, and developing the terms, we obtain a Taylor series expansion for $e^{S(x)}$, *i.e.*,

$$e^{S(x)} = \sum_{i=1}^{\infty} \alpha_i x^i,$$

where $\alpha_i$ is the sum over the coefficients of terms $x^{n_1} x^{2n_2} \cdots x^{Nn_N}$, such that $n_1 + 2n_2 + \ldots + in_i = i$. As these coefficients are equal to $\frac{a(1)^{n_1} \ldots a(N)^{n_N}}{n_1! \ldots n_N!}$, we have $c_i = \alpha_i$, which proves the first assertion of the theorem.

Next we prove the third assertion. Using the definition of $g(x)$, we have $g'(x) = g(x)S(x)$. If we look at the Taylor series of $g(x)$ we obtain

$$g'(x) = \sum_{n=0}^{\infty} (n+1)c_{n+1} x^n$$

and

$$g(x)S(x) = \left( \sum_{n=0}^{\infty} c_n x^n \right) \left( \sum_{m=0}^{\infty} a(m) x^m \right).$$

By expanding the last term, the recurrence equation in the third assertion results.

Finally, the second assertion is a classical result in analysis of convergence of series.

As $g'(x) = S'(x)g(x)$, the convergence region for the series $S(x)$ and $g(x)$ are the same, *i.e.*, $D_g = D_S$. Now, when the limit $R_S = \lim_{n\to\infty} \frac{a_n}{a_{n+1}}$ exists, the convergence region of the power series $S(x) = \sum a_n x^n$ is defined as $D_S = \{x|\ |x| < R_S\}$. But $D_g = D_S$, meaning that if $R_g = \lim_{n\to\infty} \frac{c_n}{c_{n+1}}$ exists, we have $R_S = R_g$. Now we just need to prove that the existence of $\lim_{n\to\infty} \frac{a_n}{a_{n+1}}$ implies the existence of $\lim_{n\to\infty} \frac{c_n}{c_{n+1}}$. This last property is proven in [16]. □

The above theorem gives a simple and efficient method to derive the value of the coefficients $c_N$, which are used to compute the distribution of cluster size based on Thm. 3. The method can be summarized as:

1. Write the series $S(x)$ and obtain the function that it converges to.
2. By deriving the function $g(x) = e^{S(x)}$, find a recurrence equation for the coefficient $c_n$.
3. Using the recurrence equation, the values $c_n$ are obtained, yielding the distribution of cluster sizes.

We thus have a complete theoretical characterization of the distribution of cluster size for finite values of the number of nodes.

## 2.2 Mean field analysis

The above analysis allows us to derive an analytic description of merge–split systems with finite number of nodes. However, the exact derivation does not provide much analytic insight into the large-scale behavior of systems of mobile nodes. Moreover, the proposed derivation, even if straightforward, becomes imprecise when the number of nodes $N$ becomes large because the values of $c_n$ grow almost exponentially. For as few as 70 nodes, some values of $c_n$ come close to $10^{71}$. Since the recurrence relation requires repeated summation of large values, numerical errors propagate to all values of $c_n$. For this reason, an approximation that is more amenable to calculation for systems with more than 70 nodes would be desirable. In order to deal with these two issues, we present an asymptotic analysis of merge–split processes, *i.e.*, the limit process when the number of nodes grows, $N \to \infty$. The asymptotic behavior of $\nu_N(k)$ is obtained through the next theorem:

THEOREM 5. *Suppose that*

$$R = \lim_{n\to\infty} \frac{c_n}{c_{n+1}},$$

*then for fixed $k$, we have:*

$$\lim_{N\to\infty} \nu_N(k) = a(k)R^k, k = 1, 2, \ldots$$
$$\lim_{N\to\infty} \varsigma_N(k, l) = 0, k \neq l, k, l = 1, 2, \ldots$$
$$\lim_{N\to\infty} \sigma_N^2(k) = a(k)R^k, k = 1, 2, \ldots$$

PROOF. The proof comes straight from Thm. 3 and the following observation:

$$\lim_{N\to\infty} \frac{C_{N-k}}{C_N} = \lim_{N\to\infty} \frac{C_{N-k}}{C_{N-k+1}} \frac{C_{N-k+1}}{C_{N-k+2}} \cdots \frac{C_{N-1}}{C_N} = R^k$$

□

The above asymptotic behavior provides a very interesting insight: it proves the existence of a limit when the number of nodes diverges, moreover it proves that the correlation between the number of clusters of different sizes vanishes with increasing number of nodes. This last property is called the propagation of chaos in the literature as it means that the correlation between states disappears as the number of nodes diverges.

Recently, [3] proposed a methodical approach to derive the mean field of a special class of processes where the intensity of reactions in a system vanishes with increasing number of interacting nodes, $N$. For such processes, the state occupation measure of the process converges to a mean field limit that is given as the solution of the drift equation of the state occupation measure. Unfortunately, the intensity of the above merge–split process is not vanishing as we assume that asymptotically the density of merge and split reactions converge to $K(k, l) > 0$ and $F(k|l) > 0$, respectively. Therefore, the framework from [3] is not applicable and we have to derive the mean field directly. Theorem 5 and the propagation of chaos are the basis from which the mean field will be obtained, as we will describe next.

The above theorem is difficult to apply to systems with finite number of nodes because the physical condition of the system changes as the number of nodes increases. One has to take care to ensure that the density of nodes remains constant when the number of nodes diverges. Assume that a finite system with $n$ nodes is evolving in a unit volume. In order to maintain the same physical conditions, we have to let $V(n)$ grow along with the number of nodes $n$ diverging, but ensuring that the node density remains constant and equal to $N$, *i.e.*, we are analyzing $\eta_N(k)$, the density of clusters of size $k$, when the density of nodes is equal to $N$:

$$\lim_{n\to\infty, n=N.V(n)} \nu_n(k) = \eta_N(k),$$

where the constraint $n = N.V(n)$ results from the node density being set to $N$, as indicated by subscript $N$. By extension $\eta(k,t)$ is defined as the density at time $t$ of clusters of size $k$. We also define the merge rate per volume unit as

$$K_N(k,l) = \lim_{n\to\infty, n=N.V(n)} \frac{\psi_n(k,l)}{V(n)}.$$

Similarly we define the split rate per volume unit as

$$F_N(k|l) = \lim_{n\to\infty, n=N.V(n)} \frac{\phi_n(k|l)}{V(n)}.$$

Again, we will drop index $N$ when it is obvious from the context.

We assume that the initial state $\eta(m,0),\ m = 1,\ldots,\infty$, satisfies the node conservation condition, *i.e.*, $\sum_{k\geq 1} k\eta(k,0) = N$, with $N$ denoting the density of nodes. Using the above notation and with Thm. 5, stating that the correlation between the number of clusters $\varsigma_N(k,l) \to 0$ vanishes as $N \to \infty$, we write the Kolmogorov forward equation of the Markov chain governing the merge–split process with infinite number of nodes:

$$\begin{aligned}\frac{\partial\eta(m,t)}{\partial t} &= \frac{1}{2}\sum_{l=1}^{m-1} K(l,m-l)\cdot\eta(l,t)\eta(m-l,t) && (i)\\ &- \sum_{l=1}^{\infty} K(m,l)\cdot\eta(m,t)\eta(l,t) && (ii)\\ &+ \sum_{k=m+1}^{\infty} F(k|m)\cdot\eta(k,t) && (iii)\\ &- \frac{1}{2}\sum_{l=1}^{m-1} F(m|l)\cdot\eta(m,t). && (iv) \quad (5)\end{aligned}$$

The above equation is the drift equation of the limit state variable $\eta(i,t)$; it has four terms, two from merge and two from split reactions: (i) drift into state $(m,t)$, resulting from merges between clusters of size $l$ and $m-l$ (the factor $\frac{1}{2}$ results from the symmetry of merges between clusters of size $l$ and $m-l$); (ii) drift from state $(m,t)$, resulting from merges between clusters of size $m$ and any other size; (iii) drift resulting from splits of clusters of size greater than $m$ into one cluster of size $m$ and another cluster of any other size; (iv) drift resulting from clusters of size $m$ splitting into smaller clusters (again, the factor $\frac{1}{2}$ results from the symmetry of splits of clusters of size $l$ into clusters of size $m$ and $l-m$).

As in the discrete case, there are two important variants. Without the split reaction, *i.e.*, $F(k|l) = 0$ for all $k < l$, we have a purely coalescent process described by the Smoluchowski equation [34]. Discarding the merge reaction, *i.e.*, $K(k,l) = 0$ for all $k,l$, results in a purely branching process. The Smoluchowski equation attracted an important historical interest in statistical physics [13] as it fits many real world problems, *e.g.*, polymer synthesis in chemistry, aerosol formation in the atmosphere, or phase separation in liquid mixtures. More recently, through the seminal survey by D. J. Aldous [2] the problem gained a lot of interest from the mathematical community.

Now let us assume that cluster sizes are continuous, *i.e.*, $\upsilon(x,t)$ denotes the density of clusters of size $x$ at time $t$, $x$ continuous. Then the Kolmogorov forward equation (5) turns into an integro-differential equation:

$$\begin{aligned}\frac{\partial\upsilon(x,t)}{\partial t} &= \frac{1}{2}\int_0^x K(y,x-y)\upsilon(x-y,t)\upsilon(y,t)dy\\ &- \int_0^\infty K(x,y)\upsilon(x,t)\upsilon(y,t)dy\\ &+ \int_0^\infty F(x+y|y)\cdot\upsilon(x+y,t)dy\\ &- \frac{1}{2}\int_0^x F(x|y)\upsilon(y,t)dy. \qquad (6)\end{aligned}$$

In the sequel, we are interested in deriving, when it exists, the asymptotic value $\upsilon(x) = \lim_{t\to\infty}\upsilon(x,t)$. When such an asymptotic value exists, it is the mean field approximation of the stationary distribution.

Fortunately, when the process is reversible, the stationary solution of the above integro-differential equation has a simple form that is given in the next theorem.

THEOREM 6. *The unique stationary solution*

$$\upsilon(x) = \upsilon(x,\infty)$$

*of* (6) *for a reversible Markov chain satisfying the node conservation condition is:*

$$\upsilon(x) = a(x)e^{\lambda x}, \qquad (7)$$

*where $\lambda$ is obtained subject to the node conservation condition* (1)*,*

$$\sum_{k=1}^{\infty} k\upsilon(k) = N, \qquad (8)$$

*where $N$ is the node density.*

PROOF. For (6) to reach an equilibrium, we need that $\frac{\partial\upsilon(x,t)}{\partial t} = 0$, *i.e.*,

$$\begin{aligned}&\frac{1}{2}\int_0^x K(y,x-y)\upsilon(x-y,t)\upsilon(y,t)dy\\ &\qquad + \int_0^\infty F(x+y|y)\upsilon(x+y,t)dy\\ &= \int_0^\infty K(x,y)\upsilon(x,t)\upsilon(y,t)dy + \frac{1}{2}\int_0^x F(x|y)\upsilon(x,t)dy\end{aligned}$$

Indeed, $\upsilon(x) = a(x)e^{-\lambda x}$ holds under the above condition. □

**Remark** A sufficient condition for a function $f(x)$ to be the equilibrium solution of (6) is that it satisfies simultaneously the two below equations:

$$\begin{aligned}\int_0^x K(y,x-y)f(x-y)f(y)dy &= \int_0^x F(x|y)f(x)dy,\\ \int_0^\infty K(x,y)f(x)f(y)dy &= \int_0^\infty F(x+y|y)f(y)dy.\end{aligned}$$

These two conditions are satisfied simultaneously if we require the below deterministic balance equation:

$$K(x,y)f(x)f(y) = F(x+y|y)f(x+y),\ x,y \geq 0. \qquad (9)$$

Note that a merge and split rate satisfying the reversibility conditions defined in Thm. 1 satisfies the above condition.

The next theorem shows the convergence of the system of $N$ nodes to the mean field represented by the stationary solution given in (7) for a large class of merge–split processes.

THEOREM 7. *For all functions $a(x)$ satisfying $a(x) \sim x^{\alpha}e^{\gamma x}$ when $x \to \infty$, we have:*

$$\lim_{N\to\infty} \frac{\nu_N(k)}{\upsilon(k)} = 1.$$

PROOF. Using the terms in theorem 5, we have:

$$\lim_{N\to\infty} \frac{\nu_N(k)}{\upsilon(k)} = R^k \lim_{N\to\infty} e^{-k\lambda(N)},\ k = 1, 2, \ldots$$

We proved in Thm. 4, that $R = \lim_{N\to\infty} \frac{a_n}{a_{n+1}} = e^{-\gamma}$, following the assumption of the theorem. Moreover, the node conservation condition leads to:

$$\lim_{N\to\infty} \lambda(N) = \gamma,$$

which is necessary to ensure that $\int_0^{\infty} a(x)e^{\lambda(N)x}dx = N \to \infty$. Finally, putting back in the term for $\lim_{N\to\infty} \frac{\nu_N(k)}{\upsilon(k)}$, the convergence to 1 is proven. □

This theorem yields a surprisingly simple large-scale behavior of the merge–split process that is called Mean Field Approximation (MFA). The MFA is of major interest as it provides a closed-form formula describing the cluster size behavior; in particular, it relates $a(x)$, a microscopic parameter of the merge–split process, and through it the intensity ratio $q(x, y)$, to a macroscopic property of this process, the cluster size distribution $\upsilon(k)$. This closed-form function gives insight into the properties of the cluster size distribution that cannot be inferred easily by observing the exact distribution $\nu_N(k)$. In particular the MFA shows that the head of the distribution is controlled by $a(x)$, but the tail is determined by the exponents $\gamma$ and $\lambda$, thus depending on the(finite) number of nodes.

Nonetheless, note that the convergence to the MFA is asymptotic. In particular, for large $k(N) < N$, the convergence of $\nu_N(k)$ to $a(k)R^k$ is known to be slow; *i.e.*, $\nu_N(k)$ and $\upsilon(k)$ might differ considerably for large $k(N) < N$.

### 2.2.1 Emergence of giant components

Giant components are clusters that contain a large proportion of nodes. When the number of nodes diverges, giant components become clusters of infinite size. The emergence of giant components has important practical implications: during the time when such components exist, messages can be exchanged between a fraction of nodes via a connected path. This means that the emergence of giant components is desirable and calculating the likelihood of their existence is of practical interest. It was argued before than the MFA loses its precision for large cluster sizes, but this is precisely the part of the distribution that is interesting for the analysis of giant components.

Let $L_N(\alpha)$ be defined as the number of clusters of size larger than $\alpha N$ $(0 < \alpha < 1)$ for a scenario with $N$ nodes. Then this value is derived as:

$$L_N(\alpha) = \sum_{k=\alpha N}^{N} \nu_N(k).$$

Substituting the term in Thm. 5 for $\nu_N(k)$, we obtain:

$$L_N(\alpha) = c_N^{-1} \sum_{k=\alpha N}^{N} a(k)c_{N-k}.$$

The existence of a giant component requires that

$$\lim_{N\to\infty} L_N(\alpha) > 0$$

for $\alpha \leq \alpha_0$.

Frequently, the emergence of giant components comes along with violation of the node conservation condition [14]. The physical phenomenon corresponding to the emergence of an infinite-size cluster and violation of node conservation is called gelation or precipitation. Before gelation, the node conservation condition holds; after gelation, $\sum_{k=0}^{\infty} k\nu_{\infty}(k)$ decreases as the number of nodes available in the non-gelated phase decreases. This is a well-known phase transition phenomenon in chemistry, where all (or a large proportion of) particles in a suspension evolve from a fluid phase into a semi-solid (gelation) or solid (precipitation) phase. As our interest in this paper is the stationary distribution that emerges after a long time, the resulting stationary distribution might not satisfy the node conservation condition. In particular, this means that some proportion of the nodes are not participating in the merge–split process because they are "captured" in a giant component, or because they have been withdrawn from the interaction medium.

In the context of this work we are more interested in systems with finite size. For these systems we can ensure node conservation by setting the parameters of the function $a(x)$ appropriately. We illustrate this procedure further in Sec. 2.3.

### 2.2.2 Case study

To show the convergence to the mean field and illustrate the above effects, we study two cases: $a(i) = \beta$ and $a(i) = \frac{\beta}{i}$.

**Case 1:** $a(i) = \beta$**.** This is the case where the merge and split rates are constant and $q(i, j) = \frac{1}{\beta}$. The function $S(x)$ is derived as

$$S(x) = \sum_{i=1}^{\infty} \beta x^i = \frac{\beta x}{1 - x},$$

with $D_S = (-1, 1)$ and $g(x) = e^{\frac{\beta x}{1-x}}$. By deriving $g(x)$ we have $(1 - x^2)g'(x) = \beta g(x)$, yielding the following recurrence equations for $n = 1, 2, \ldots$:

$$c_0 = 1, c_1 = \beta;$$

$$(n + 1)c_{n+1} = (2n + \beta)c_n - (n - 1)c_{n-1},$$

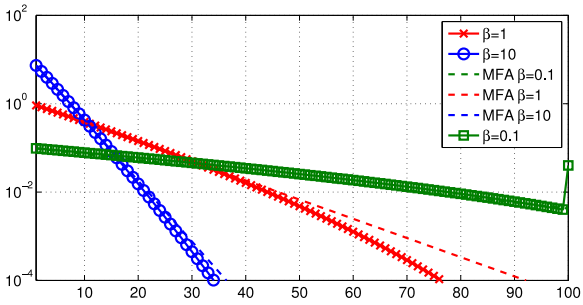
which leads to a monotonically increasing sequence $c_n,\ n > 0$. Therefore $\nu_N(k)$ is monotonically decreasing with $k$, $1 \leq k < N$ (it might increase for $k = N$).

Applying the mean field formula given in Thm. 6, with $a(x) = \beta$ we obtain $\lambda(N) = -\sqrt{\frac{\beta}{N}}$:

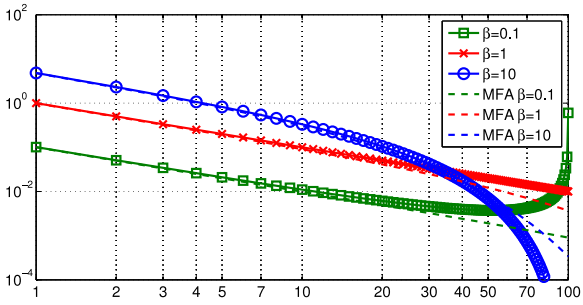
$$\upsilon(x) = \beta e^{-\sqrt{\frac{\beta}{N}}k}, \tag{10}$$

showing an exponential decrease of the number of clusters with the cluster size. The asymptotic distribution predicted by Thm. 5 is derived by noting that $R = \lim_{k\to\infty} \frac{a_k}{a_{k+1}} = 1$ and $\lim_{N\to\infty} \nu_N(k) = \beta$. Moreover $\lim_{N\to\infty} \lambda(N) = 0$, showing that $\lim_{N\to\infty} \upsilon_N(k) = \lim_{N\to\infty} \nu_N(k) = \beta$.

In Fig. 1a we show the distribution of cluster sizes obtained by the method described for finite number of nodes for a 100 nodes scenario, as well as the MFA given in Thm. 6. This demonstrates the remarkable quality of the mean field approximation, at least for small values of cluster sizes. Note that Fig. 1a also shows the loss of precision of the MFA for large $k$ for finite numbers of nodes, $N$. These observations are in line with the theoretical analysis that predicted the MFA to be looser for large cluster sizes.




(a) $a(i) = \beta$; note the logarithmic $y$ axis.




(b) $a(i) = \frac{\beta}{i}$; note the logarithmic $x$ and $y$ axes.

**Figure 1: Cluster size vector $\nu_{100}(k)$: exact derivation compared to result of MFA; $N = 100, \beta \in \{0.1, 1, 10\}$.**

**Case 2: $a(i) = \frac{\beta}{i}$.** In this case $q(i,j) = \frac{ij}{\beta(i+j)}$. Such a function $a(i)$ can be used when clusters merge with a rate proportional to the product of their size and split with a rate proportional to their size. With this assumption, the function $S(x)$ is derived as

$$S(x) = \beta \sum_{i=1}^{\infty} \frac{x^i}{i} = -\beta \log(1-x), \; D_S = (-1, 1).$$

Consequently $g(x) = \frac{1}{(1-x)^\beta}$. This results in

$$\begin{aligned} c_n &= \frac{\beta(\beta+1)\dots(\beta+n-1)}{n!} \\ &= \frac{\Gamma(n+\beta)}{\Gamma(\beta)\Gamma(n+1)}, \; n = 0, 1, \dots \end{aligned} \tag{11}$$

Applying Thm. 3 generates the statistics of cluster sizes in a straightforward way. In particular for $\beta = 1$ we have $c_n = 1$ resulting in $\nu_N(k) = \frac{\beta}{k}$, which is independent of $N$.

The MFA for $a(x) = \frac{\beta}{x}$ is obtained as

$$\upsilon(x) = \frac{\beta}{x} e^{-\frac{\beta}{N}x} \tag{12}$$

and the asymptotic distribution predicted by Thm. 5 then becomes $\lim_{N\to\infty} \nu_N(k) = \frac{\beta}{k}$.

We show in Fig. 1b the distribution of cluster sizes obtained for a 100 nodes scenario with $a(i) = \frac{\beta}{i}$ as well as the relevant MFA. Here also the MFA results in a remarkable approximation for small to moderate values of cluster sizes. However the approximation becomes looser for large cluster size because of the accumulation effect of finite $N$. By comparing Figs. 1a and 1b, it can be seen that large size clusters are more frequent with $a(i) = \frac{\beta}{i}$ than when $a(i) = \beta$. In particular, in Fig. 1b the exact derivation of the cluster size vector for $\beta = 0.1$ yields 0.60 clusters with size 100, implying that only 40% of the nodes are in smaller clusters, hence the distribution is concentrated on a single cluster with 100 nodes.

Analysis of the correlation structure of the number of clusters gives interesting insights for this case. We show in Fig. 2, the correlation coefficient $\frac{\varsigma_N(k,l)}{\sigma_N(k)\sigma_N(l)}$ obtained through Thm. 3 for different values of $\beta$ when $a(i) = \frac{\beta}{i}$. For $\beta \leq 0.1$, we observe a relatively strong correlation between values $\nu_N(k)$ and $\nu_N(N-k)$ (the values on the antidiagonal). Moreover, there is also a strong correlation between $\nu_N(N)$ and all other $\nu_N(k)$ (last row and column of the correlation coefficient matrix). This means that there are frequent direct transition from clusters of size $k < N$ to cluster of size $N$. The correlation on the antidiagonal can be interpreted as resulting from this last fact; most transitions are $C_N \to C_{N-k} + C_k$ and $C_{N-k} + C_k \to C_N$, *i.e.*, the number of clusters of size $k$ and $N - k$ are expected to be almost equal and this is confirmed by observing the curves in Fig. 1b that shows an almost symmetric curve of $\nu_N(k)$. When $\beta$ becomes closer to 1, other transitions also appear. Nevertheless, for values of $\beta < 1$, these reactions occur almost exclusively for clusters of large size.

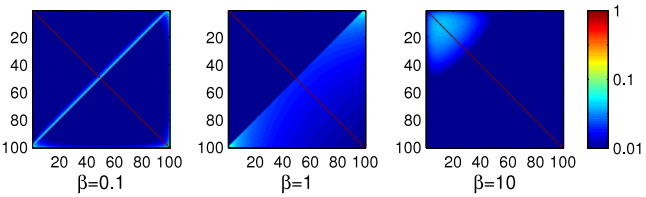



**Figure 2: Correlation coefficient between number of clusters $\frac{\varsigma_{100}(k,l)}{\sigma_{100}(k)\sigma_{100}(l)}$ for 100 nodes obtained for $a(i) = \frac{\beta}{i}$ for different values of $\beta$, plotted with a logarithmic scale**

For $\beta > 1$, the correlation structure changes and becomes concentrated on the upper left triangle, representing small cluster sizes, which can be interpreted by observing that now most transitions involve clusters of small size and only rarely large clusters.

## 2.3 Empirical fitting and effect of parameters

In practice we normally have access to microscopic information about the merge and split rate that results from the particular mobility pattern of a scenario. From these information one can estimate the intensity ratio $\hat{q}(i,j)$, that can be fitted to any functional form. However we saw previously that in order for the Markov process defined by the merge–split reactions to be reversible we should be able to find a function $a(i)$ such that $q(i,j) = \frac{a(i+j)}{a(i)a(j)}$. Moreover, Thm. 7, proving the convergence to the mean field, suggests an asymptotic convergence to $a(i) = \beta \frac{e^{\gamma i}}{i^\alpha}$. Using such a functional form for $a(i)$ results in

$$q(i,j) = \frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha},$$

showing that $q(i,j)$ does not depend on $\gamma$. Therefore $\alpha$ and $\beta$ should be derived by fitting to the empirical intensity ratio $q(i,j)$; $\gamma$ can be estimated by applying the node conservation condition (1) to the exact cluster size distribution $\nu_N(k)$, *i.e.*, $\gamma$ is chosen such that the resulting distribution $\nu_N(k)$ satisfies (1).

The parameters $\alpha$ and $\beta$ are derived by fitting empirically derived values of intensity ratios to a function $\frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha}$ by a non-linear least-mean-square (LMS) technique. Frequently, the number of observed merge and split events becomes very small in particular for large cluster sizes, reducing their statistical value. A weighting equal to $\sqrt{m(i,j)s(i,j)}$ (where $m(i,j)$ is the number of merge events observed between clusters of size $i$ and $j$ and $s(i,j)$ is the number of split events of clusters of size $i+j$ to two clusters of sizes $i$ and

$j$, respectively) is applied to every measured intensity ratio. Moreover, for some cases the dynamical range of measured intensity ratio $q(i,j)$, is very large, *e.g.*, for small $i$ and $j$, $q(i,j) \sim 0.001$ and for large $i$ and $j$, $q(i,j) \sim 10$. In such cases we calibrate $\log q(i,j)$ to $\log \frac{i^\alpha j^\alpha}{\beta(i+j)^\alpha}$.

Knowing $\alpha$, $\beta$ and $\gamma$, the exponent $\lambda$ to be used in the MFA can be obtained by solving the following equation:

$$\beta \sum_{k=1}^{N} \frac{e^{(\gamma+\lambda)k}}{k^{\alpha-1}} = N. \quad (13)$$

Solving this last equation provides all the information needed for predicting the cluster size distribution.

### 2.3.1 *Impact of parameters* $\alpha, \beta, \lambda, \gamma$

The expressions used in the above parameter fitting procedure provide insight as to the influence of the parameters on the shape of the cluster size distribution; this curve has two distinct parts: its head and its tail. The head of the distribution contains two essential pieces of information: the number of isolated nodes, *i.e.*, nodes that are not connected to any other node, and the slope of the decay of the distribution. The mean field approximation yields that for small values of cluster sizes, the distribution can be approximated as a polynomial with exponent $-\alpha$, and the number of isolated nodes can be estimated as equal to $\upsilon(1) = \beta e^{\lambda}$, where, in contrast to $\alpha$ and $\beta$, $\lambda$ depends on the number of nodes.

The tail of the distribution is governed partly by the number of nodes $N$, which directly controls $\gamma$ and $\lambda$. If $N$ increases, the exponent $\gamma + \lambda$ turns positive, yielding a greater number of large clusters; indeed in that case, one may expect a bump at the tail of the distribution. This bump is the sign for the emergence of giant components, as we will discuss in the next section.

### 2.3.2 *Emergence of giant components*

Giant components are a well-known phenomenon in percolation theory [12]. In our case, the conditions under which such giant components emerge are determined by the value of $\lambda + \gamma$, which is controlled by (13). This equation states that the exponent $\lambda + \gamma$ is positive if and only if $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}} < N$. For example if $\alpha = 1$, $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}} = N$, yielding two qualitative behaviors, depending on the value of $\beta$. For $\beta < 1$, one may observe giant components, whereas for $\beta \geq 1$ no such components emerge. This is in line with the analysis given in Sec. 2.2.2, where $\beta = 1$ was found to be a boundary value for two distinctive behaviors of the correlation structure of the finite system of nodes [1]. Indeed, the smaller the value of $\beta \sum_{k=1}^{N} \frac{1}{k^{\alpha-1}}$ (*i.e.*, the larger $\alpha$ and the smaller $\beta$), the larger the exponent $\lambda + \gamma$ and the stronger the tail bump, implying a larger giant component.

This last property gives intuition as to the impact of the parameters and helps to interpret them in terms of mobile network scenarios. A large $\alpha$ and a small $\beta$ relative to $N$, *i.e.*, small $\beta/N$ means that the system of $N$ nodes has one or several giant components in its stationary state, accompanied by isolated nodes that merge and split with the giant components. In contrast, a small value of $\alpha$, independently of $\beta$, implies that the network will remain an archipelago of disconnected clusters that merge and split among each other.

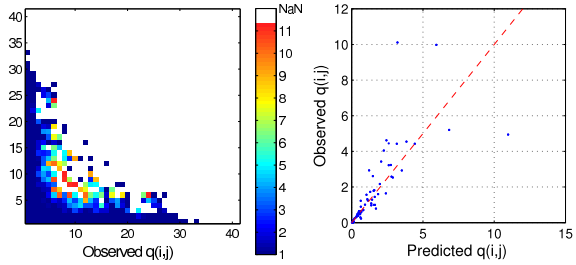
## 3. VALIDATION

[1]A similar analysis, with more rigorous analytical basis, can be done for the exponent $\gamma$ alone (in place of $\lambda + \gamma$) and leads to mathematically stronger results; this analysis is omitted due to lack of space

To this point, the analysis provided was strictly of theoretical nature. In this section we aim to validate that this mathematical analysis is of practical interest for predicting the behavior of realistic mobile wireless networks. We will do this by analyzing a variety of scenarios: three real world scenarios as well as a synthetic random walk scenario. First, we will use the contact trace from Infocom 2005 as an example of a realistic mobile network and show that it can be described by a merge–split model. In the second part we study the random walk simulation, which serves to relate scenario parameters such as node density to the parameters of the merge-split process. Finally, we will analyze two large-scale traces based on GPS position records from taxis in San Francisco and Shanghai to show the applicability of our model to scenarios with hundreds and thousands of nodes.

## 3.1 Infocom 2005 contact trace

In this subsection, we study the scenario described in [7]. In this experiment, 41 conference attendees of Infocom 2005 carried a small Bluetooth contact logger during the three days of the conference. Based on the Bluetooth contacts logged as tuples {device hardware address, contact start time, contact end time}, the connectivity graph has been reconstructed, allowing the merge and split rate function to be estimated empirically and their intensity ratio (defined in (2)) to be derived. We plot the ratio $q(i,j)$ of those values in Fig. 3a: clearly, $q(i,j)$ increases with cluster size; nonetheless, a large part of the rate function remains undefined (shown with white color corresponding to "not a number" (NaN) in the figure) as no merge and split involving these values has been observed.



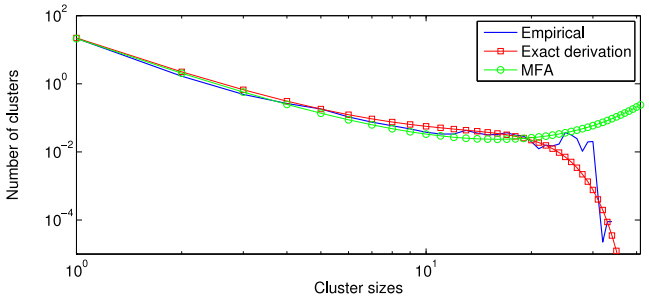

(a) Observed merge/split ratio $q(i,j)$

(b) Predicted $\hat{q}(i,j)$ *vs.* observed ratio $q(i,j)$

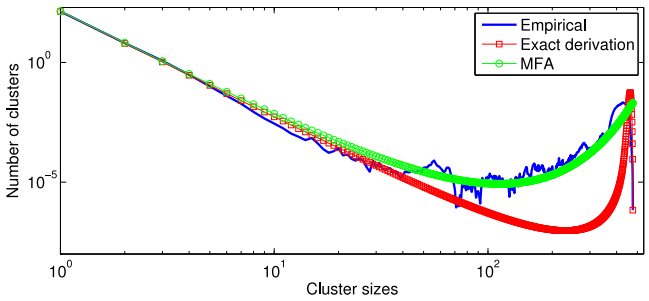
**Figure 3: Merge/split ratio $q(i,j)$ and comparison with prediction in the Infocom scenario**

Applying the weighted least-mean-squares fitting described in Sec. 2.3 to the measured intensity ratio yields an estimation of $\hat{\alpha} = 3.71 \pm 0.1$, $\hat{\beta} = 16.73 \pm 0.95$ with a remarkable $R^2 = 0.998$ goodness of fit indicator. The value $\hat{\gamma} = 0.83$ is obtained by enforcing node conservation on the distribution $\nu_N(k)$. By enforcing node conservation on the MFA, one can derive $\lambda = -0.66$, resulting in $\lambda + \gamma = 0.17$ and therefore the emergence of a giant component. This can be verified by noting that $16.73 \sum_{k=1}^{41} \frac{1}{k^{3.71}} = 21.25 < 41$. In Fig. 3b, we plot the observed ratio $q(i,j)$ against the predicted ratio $\hat{q}(i,j) = \frac{a(i+j)}{a(i)a(j)}$, with $a(x) = \frac{16.73}{x^{3.71}e^{0.86x}}$.

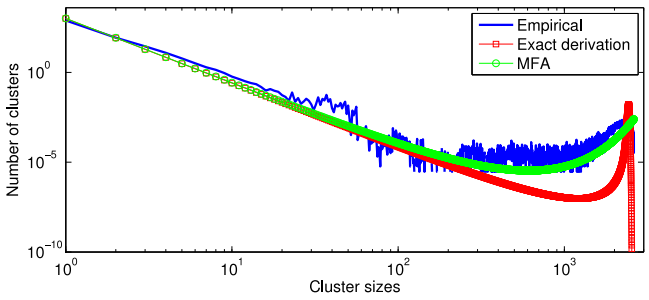
In Fig. 4a, we compare the cluster size vector observed over the entire trace with the distribution predictions introduced previously, *i.e.*, the exact derivation from Sec. 2.1, and the MFA from Sec. 2.2. The two distributions predict the empirical distribution with remarkable accuracy and the difference between the MFA and the exact derivation are in line with the analysis provided previously.




(a) Infocom




(b) San Francisco taxicabs




(c) Shanghai taxicabs

**Figure 4: Empirical cluster size vector for real-world scenarios with exact derivation and MFA**

Note that even though the number of nodes is quite small, the exact derivation still yields a good prediction of the cluster size vector.

## 3.2 Synthetic random walk scenario

As a second scenario to validate our approach we used a synthetic random walk scenario. For this purpose, we run an extensive set of simulations with a simple home-grown mobility simulator that models mobile nodes moving according to the following random walk mobility model: at initial time $t = 0$, $N = 1000$ nodes are placed uniformly at random in a square area with variable side length. Every node is assigned a random direction in $[0, 2\pi)$. All nodes move in the assigned direction for one distance unit, then they pick a new direction at random. If the trajectory of a node leads outside the simulation area it is reflected at the closest border. A link between two nodes is up if their Euclidean distance is less than the transmission range $r_{TX} = 100$. In Fig. 5, we plot the fitted values of $\alpha$ and $\beta$ as a function of coverage (defined as the ratio between the area covered by the aggregated transmission range of all nodes and the simulation area). Since the coverage decreases with the square of the area side length, we picked values with ratio $1 : \sqrt[4]{2}$. The coverage values range from 0.44 (corresponding to an area side length of 8450) up to 12 (side length 1618). We

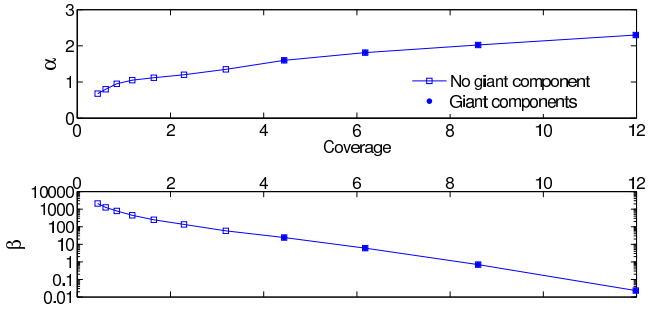
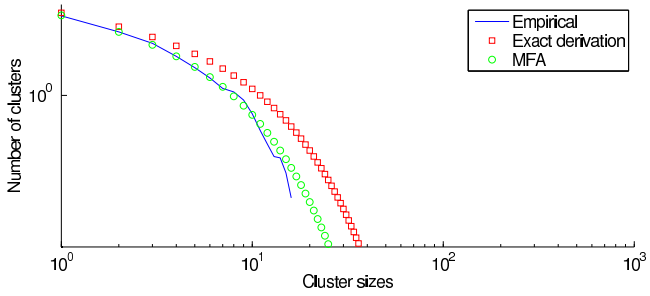
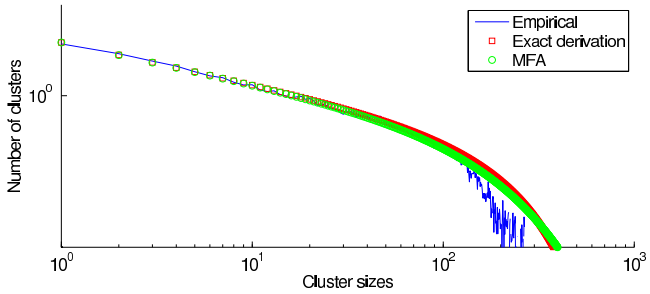



**Figure 5: Estimated values of parameters $\alpha$ and $\beta$ as a function of coverage in the random walk scenario**

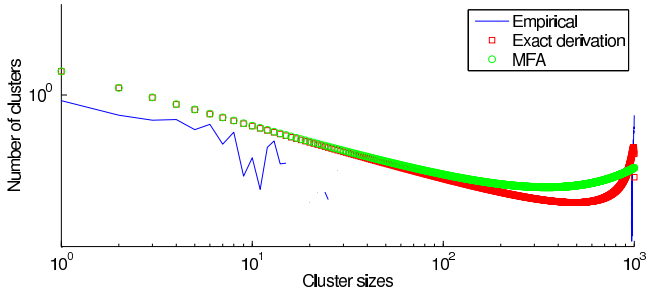
observe that with increasing coverage, $\alpha$ increases almost linearly and $\beta$ decreases almost exponentially. Nevertheless, this holds only for coverage values greater than 4.4, where giant components may emerge. The figure also indicates the value of $\alpha$ at which large clusters can be expected to emerge. To further illustrate the emergence




(a) Coverage 0.8




(b) Coverage 3.2




(c) Coverage 8.6

**Figure 6: Empirical cluster size vector for random walk scenario with exact derivation and MFA**

of giant components, we plot the empirical cluster size vector as well as the cluster size vector calculated according to the exact and

the mean field derivation from Sec. 2 in Fig. 6 for three coverage values. While no giant component emerges for coverage values of 0.8 or 3.2, at coverage 8.6, the empirical result as well as the exact and the mean field prediction indicate the existence of giant components. Note that since in this scenario node distribution is uniform, connectivity, *i.e.*, the emergence of a single giant component involving all nodes, requires coverage of about 12. At this high node density, almost no merge and split events are observed and 99% of time all nodes are in the same cluster, corresponding to the network being connected. This is in line with results of continuum percolation theory [12].

## 3.3 Taxicab mobility traces

We draw further statistics from two mobility traces based on GPS (Global Positioning System) position records from taxicabs. Since our model is based on cluster merge and split events, the raw position reports from GPS traces need to be translated to node adjacency matrices. In line with other recent publications [35, 8], we define two nodes to be connected if their distance does not exceed a fixed transmission distance (we use 200 meters, roughly corresponding to the typical range of IEEE 802.11). Note that under this simple definition of a link, a link being up between two nodes indicates only that those nodes are close enough for communication to be feasible in principle, but it does not imply that communication would succeed in practice. (In contrast, in the Infocom contact trace, two nodes being connected means that they actually exchanged data.) Nonetheless, since the calibration (*cf.* Sec. 2.3) incorporates the propagation model in the same way as the observed cluster size distribution against which we validate the prediction, our validation remains valid.

While one could emulate the effects of wireless propagation post facto, we expect that it would not significantly affect the quality of our prediction.

As the GPS position reports contain outliers, we use an MAD (Median of the Absolute Deviation) filtering procedure [11] on the raw positions. We consider only reports that are no farther apart than certain distance and time limits and interpolate positions between reports to increase temporal resolution to ten seconds. Finally, to reduce the impact of daily patterns, we use only the time range 8AM until 12PM.

### *3.3.1 San Francisco taxicab mobility trace*

The traces from the San Francisco Cabspotting project have previously been studied in the context of DTN [30]; our trace contains 11.2 million GPS positions from 517 cabs over the course of 21 days. We applied the model calibration over this data set and obtained estimates of $\hat{\alpha} = 4.437 \pm 0.004$ and $\hat{\beta} = 133.2 \pm 0.4$, $\hat{\gamma} = 0.3648$ with an $R^2 = 0.995$. The MFA was calibrated with a value $\hat{\lambda} = -0.3258$. The comparison of the empirical distribution of cluster sizes over the San Francisco taxis and the comparison with the exact derivation and the MFA are shown in Fig. 4b. This figure shows good agreement between the empirical distribution and the MFA. However, the quality of the prediction at the tail of the exact distribution degrades. This is to be expected as the number of nodes (being much larger than 70) yields $c_n$ values beyond the limit of double precision floating point arithmetic. Indeed, for large scenarios the MFA can be the more suitable approximation as the figure shows. Of note, the San Francisco trace yields a larger value of $\alpha$ than the simulation scenario (Sec. 3.2), showing that in real scenarios nodes have a higher tendency to gather and build large clusters. Interestingly this tendency is even higher than for the Infocom scenario, where the exponent $\alpha$ is larger when the proportion of isolated nodes for the two scenarios are between 25% to 35%. This can be explained by the fact that taxis frequently gather at hot spots (train stations, restaurants, *etc*.), leading to a highly non-uniform distribution ([30] studies hot spots in this trace).

### *3.3.2 Shanghai taxicab mobility trace*

The Shanghai taxicab traces were collected by the Traffic Information Grid Team at Shanghai Jiaotong University [20]. The data consists of GPS position reports from 4063 taxis in Shanghai, spanning 28 days. We calibrate the ratio of intensity function as before and obtain $\hat{\alpha} = 3.602 \pm 0.1$ and $\hat{\beta} = 1007 \pm 3$ with an $R^2 = 0.9823$. The other parameters are also obtained as $\hat{\gamma} = 0.23$ and $\hat{\lambda} = -0.2242$. The comparison of the empirical distribution of cluster sizes over the Shanghai taxicab trace and the comparison with the exact derivation and the MFA are shown in Fig. 4c. This figure shows very good agreement between the empirical distribution and the MFA. As expected the exact derivation fails to provide a perfect prediction as the number of nodes leads to computational artefacts. For this scenario, $\alpha$ is in the order of the Infocom scenario and smaller than the San Francisco scenario. The effect of the difference in $\alpha$ can be seen by observing that the bump in the tail of the San Francisco scenario is more pronounced than the one of the Shanghai scenario. The difference between the two taxicab scenarios might come from the differences in the gathering pattern of cabs and from the geographical and topographical difference between these two cities; hot spots of this trace are studied in [23].

# 4. RELATED WORK

While routing in mobile ad hoc networks (MANETs) implicitly assumes a *connected* network and has often been studied with simulation, measurements (*e.g.*, [6, 10, 1, 24]) found the scenarios for which results could be obtained to be *disconnected*. As a result, schemes to leverage single-hop communication opportunities *(contacts)* recently attracted a lot of interest in the context of delay-tolerant networking (DTN) [21]. Studying in particular *space-time paths, i.e.*, multi-hop paths arising *over time*, Chaintreau *et al*. [9] observed a "small world" behavior in many mobility traces, as the diameter of those networks is around four to six hops.

At a more abstract level, the gathering behavior of people has inspired several mobility models and forwarding schemes which explicitly account for clustering properties (*e.g.*, [27]); algorithms building upon this property often complement MANET routing with opportunistic forwarding between clusters (*e.g.*, [31]). Piórkowski *et al*. [30] derived a heterogeneous random walk mobility model that includes clustering as a feature of the scenario, rather than as the result of social behavior of the nodes—thus rendering nodes statistically indistinguishable. Similar to their work, our model is also completely agnostic to social behavior of nodes. Focusing specifically on the *order* in which contacts happen, Tang *et al*. [33] introduced a temporal distance metric based on the concept of reachability and connected components which quantifies the efficiency and diffusion performance of social networks.

Regarding the classification of network scenarios, the phase transition of connectivity of asymptotically large networks has been studied using percolation theory already in [29]; further research analyzed more realistic scenarios [4, 32]. More recently, complex network research has analyzed the emergence of giant components in mobile scenarios. For example, Wang *et al*. [35] studied usage locations of several million mobile phone users and observed large clusters emerging in urban areas. Similarly to our comparison of real-world traces with synthetic mobility, they also observed that the empirical scenario yields a less abrupt increase of the size of the largest component as compared to a random geometric graph model. The size of the giant component has also been studied by

Hekmat *et al*. [18], where a lognormal propagation model is introduced to smooth the phase transition. From a more abstract angle, Borrel *et al*. [5] proposed a methodology for classifying networks on the entire connectivity range and presents an interesting taxonomy of such scenarios.

Finally, Chaintreau *et al*. [8] used similar methodology to analyze a disconnected network under a different aspect, namely the emergence of a spatial mean field describing the age of the latest update received by mobile nodes running a gossip protocol.

# 5. CONCLUSION AND FUTURE WORK

Beginning with the observation that real world mobile networks may comprise sizable connected components (clusters), we develop a model for predicting the distribution of the size of those clusters based on the rate at which they merge and split. This model allows capturing heterogenous node distribution as well as multi-hop paths and yields a closed-form result for finite number of nodes. We then show that with increasing number of nodes, the process converges to a mean field behavior. This means that the model yields a simple expression that translates the observable merge and split rate of a scenario to the stationary cluster size distribution. In order to validate the predictive quality of this model, we use both synthetic random walk mobility as well as three real-world mobility traces ranging from dozens to thousands of nodes. We find that the exact derivation as well as the mean field approximation predict cluster size distribution that matches the empirical distribution with remarkable accuracy.

Motivated by these results, we have continued our analysis to obtain several results on the *dynamics* of the merge–split process.

- We analyze the time to reach the stationary state, yielding further insight into the temporal characteristics of a scenario;
- for an individual node, we derive the probability distribution of its cluster's size following the subsequent merge event;
- we study the fraction of nodes that may communicate in a disconnected network given a delay bound.

These results are to be published in the near future.

We hope that our work might help to further our understanding of the complex clustering phenomena in mobile wireless networks and motivate the consideration of partial paths for forwarding algorithms [17], instead of focusing solely on either contacts (DTN) or end-to-end paths (MANET).

# 6. ACKNOWLEDGMENTS

We are grateful to our shepherd, Giovanni Pau, for his valuable advice in improving the presentation of this work, and to the anonymous reviewers for their thoughtful comments. K. Salamatian wishes to thank J.-Y. Le Boudec for insightful discussion during his sabbatical leave at EPFL in 2007 on the nature of the mean field approach and how to generalize it.

# 7. REFERENCES

[1] D. Aguayo, J. Bricket, S. Biswas, G. Judd, and R. Morris. Link-level measurements from an 802.11b mesh network. In *ACM SIGCOMM '04*.

[2] D. J. Aldous. Deterministic and stochastic models for coalescence (aggregation, coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, 5:3–48, 1997.

[3] M. Benaim and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65(11-12):823–838, 2008.

[4] L. Booth, J. Bruck, M. Franceschetti, and R. Meester. Covering algorithms, continuum percolation and the geometry of wireless networks. *Ann. Appl. Probab.*, 13(2):722–741, 2003.

[5] V. Borrel, M. H. Ammar, and E. W. Zegura. Understanding the wireless and mobile network space: a routing-centered classification. In *ACM SIGCOMM '07 CHANTS Workshop*.

[6] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva. A performance comparison of multi-hop wireless ad hoc network routing protocols. In *ACM MobiCom '98*.

[7] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *IEEE INFOCOM '06*.

[8] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: Spatial mean field regime. In *ACM SIGMETRICS '09*.

[9] A. Chaintreau, A. Mtibaa, L. Massoulie, and C. Diot. The diameter of opportunistic mobile networks. In *ACM CoNEXT '07*.

[10] K.-W. Chin, J. Judge, A. Williams, and R. Kermode. Implementation experience with MANET routing protocols. *SIGCOMM Comput. Commun. Rev.*, 32(5):49, 2002.

[11] L. Davies and U. Gather. The identification of multiple outliers. *J. Amer. Stat. Assoc.*, 88(423):782–792, 1993.

[12] O. Dousse, M. Franceschetti, N. Macris, R. Meester, and P. Thiran. Percolation in the signal to interference ratio graph. *J. Appl. Probab.*, 43(2):552–562, 2006.

[13] R. L. Drake. *A General Mathematical Survey of the Coagulation Equation*, volume 2-3 of *International reviews in aerosol physics and chemistry*, pages 201–376. Pergamon Press, 1971.

[14] P. B. Dubovskii and I. W. Stewart. Existence, uniqueness and mass conservation for the coagulation-fragmentation equation. *Math. Methods in the Appl. Sci.*, 19:571–591, 1996.

[15] V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot. Diversity of forwarding paths in pocket switched networks. In *ACM IMC '07*.

[16] G. A. Freiman and B. L. Granovsky. Asymptotic formula for a partition function of reversible coagulation-fragmentation processes. *Israel Journal of Mathematics*, 130(1):259–279, December 2002.

[17] S. Heimlicher, M. Karaliopoulos, H. Levy, and T. Spyropoulos. On leveraging partial paths in partially-connected networks. In *IEEE INFOCOM '09*.

[18] R. Hekmat and P. V. Mieghem. Connectivity in wireless ad-hoc networks with a log-normal radio model. *Mobile Networks and Applications*, 11(3), 2006.

[19] G. Holland and N. H. Vaidya. Analysis of TCP performance over mobile ad hoc networks. In *ACM MobiCom '99*.

[20] H.-Y. Huang, P.-E. Luo, M. Li, D. Li, X. Li, W. Shu, and M.-Y. Wu. Performance evaluation of SUVnet with real-time traffic data. *IEEE Trans. Vehicular Techn.*, 56(6), Nov. 2007.

[21] E. P. Jones and P. A. Ward. Routing strategies for delay-tolerant networks, 2006. `http://ccng.uwaterloo.ca/~pasward/Publications/dtn-routing-survey.pdf`.

[22] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979.

[23] J. Lee, K. Lee, J. Jung, and S. Chong. Performance evaluation of a DTN as a city-wide infrastructure network. In *ACM CFI '09*.

[24] V. Lenders, J. Wagner, S. Heimlicher, M. May, and B. Plattner. An empirical study of the impact of mobility on link failures in an 802.11 ad hoc network. *IEEE Wireless Communications*, 15(6):16–21, Dec. 2008.

[25] A. A. Lushnikov. Coagulation in finite systems. *Journal of Colloid and Interface Science*, 65(2):276 – 285, 1978.

[26] A. H. Marcus. Stochastic coalescence. *Technometrics*, 10(1):133–143, Feb. 1968.

[27] M. Musolesi and C. Mascolo. CAR: Context-Aware Adaptive Routing for delay-tolerant mobile networks. *IEEE Transactions on Mobile Computing*, 8, 2008.

[28] J. Ott, D. Kutscher, and C. Dwertmann. Integrating DTN and MANET routing. In *ACM SIGCOMM '06 CHANTS Workshop*.

[29] T. Philips, S. Panwar, and A. Tantawi. Critical connectivity phenomena in multihop radio models. *IEEE Transactions on Information Theory*, 35(5):1044–1047, 1989.

[30] M. Piórkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *COMSNETS '09*.

[31] N. Sarafijanovic-Djukic, M. Piórkowski, and M. Grossglauser. Island hopping: Efficient mobility-assisted forwarding in partitioned networks. In *IEEE SECON '06*.

[32] M. A. Serrano and M. Boguñá. Percolation and epidemic thresholds in clustered networks. *Phys. Rev. Lett.*, 97(8):088701, Aug 2006.

[33] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Characterising temporal distance and reachability in mobile and online social networks. *ACM SIGCOMM '10*.

[34] M. von Smoluchowski. Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen. *Phys. Z.*, 17:557–571 and 585–599, 1916.

[35] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding the Spreading Patterns of Mobile Phone Viruses. *Science*, 324(5930):1071–1076, 2009.