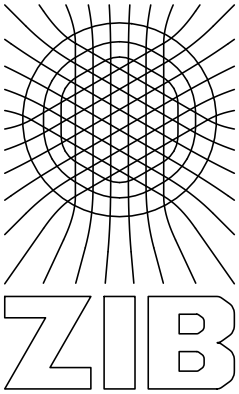
Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany



RALF BANISCH[1], CHRISTOF SCHÜTTE[1,2]
AND NATAŠA DJURDJEVAC CONRAD[1,2]

[1]*Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*

[2]*Zuse Institute Berlin, Germany*

Module Detection in Directed Real-World Networks

ZIB-Report 14-13 (April 2014)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Module Detection in Directed Real-World Networks

Ralf Banisch*
Institut für Mathematik und Informatik, FU Berlin

Christof Schütte† and Nataša Djurdjevac Conrad‡
Institut für Mathematik und Informatik, FU Berlin, and Zuse Institute Berlin (ZIB)§
(Dated: April 25, 2014)

We investigate the problem of finding modules (or clusters, communities) in directed networks. Until now, most articles on this topic have been oriented towards finding complete network partitions despite the fact that this often is unwanted. We present a novel random walk based approach for non-complete partitions of the directed network into modules in which some nodes do not belong to only one of the modules but to several or to none at all. The new random walk process is reversible even for directed networks but inherits all necessary information about directions and structure of the original network. We demonstrate the performance of the new method in application to a real-world earthquake network.

Real-world systems are often modeled as complex networks (or graphs). Understanding the behavior of these systems is thus closely related to investigating the topology and dynamics of the associated networks. One of the most challenging questions in this direction is efficient and accurate identification of network **modules**, i.e. densely inter-connected subgraphs having sparse connections to the rest of the network. During the last years, different approaches have been proposed to detect modules in networks [1–5], see [6] for an exhaustive review. However, most of them suffer from the following limitations: they (i) consider only complete partitions of a network where every node is assigned to exactly one module, and (ii) can only be applied to undirected networks. In other words, methods for *non-complete* partitions of *directed* networks are rare. For example, the most prominent technique for module finding based on modularity maximization [2] has been generalized to directed networks [5, 7] but still aims for a complete partition; the situation is true for pure graph algorithms like mincut and most other clustering techniques. The largest family of methods that allow for non-complete partitions of directed networks use the Stochastic Blockmodel [8, 9] and Variational Bayesian [10] approaches which often suffer from convergence problems.

The two main results of this paper are: (a) a generalization of the well known modularity function [2] to directed networks which encodes directional information correctly, in combination with (b) a new method for finding non-complete/fuzzy partitions of directed, weighted networks using random walk processes (RW) [11] and spectral methods because of their optimality [12]. We consider the network $G = (V; E)$, where V is the set of nodes and E the set of edges. We assume that the network is strongly connected, in particular that it has no sinks and sources. Let us denote the weight of edge $(xy) \in E$ by $K(x, y)$ and the weighted in- and out-degree of a node by $K^+(x) = \sum_y K(x, y)$, $K^-(x) = \sum_y K(y, x)$ respectively. Now, we can define a time-discrete RW process $(X_n)_n$ on the network by specifying the following transition matrix P

$$p_{xy} = \frac{K(x, y)}{K^+(x)}. \tag{1}$$

Because of our assumptions on the networks we consider, P is ergodic with unique stationary distribution π. If the network is undirected then K is symmetric and P is reversible.

A widely used approach for finding modules $C_m \subset V$ in directed networks is to optimize a quantity Q called *modularity* over all module assignment functions χ_m, where $\chi_m(x) = 1$ for $x \in C_m$ and $\chi_m(x) = 0$ for $x \notin C_m$. The modularity function has been introduced by Newman and Girvan [2, 13] for undirected, unweighted networks and various generalizations have been proposed [5, 7, 14] that extend the definition to directed networks, all of which can be written as

$$Q = \sum_m \sum_{x,y} \chi_m(x) \left[\pi_x p_{xy} - \pi_x \pi_y\right] \chi_m(y), \tag{2}$$

Optimization of Q is known to be NP-hard [15], but various fast heuristic algorithms exist [16, 17]. We will refer to this family of methods as *generalized Newman-Girvan (gNG)* approaches. It is obvious that *gNG* methods suffer from two main drawbacks. First, they consider only full partitions. Second, Q even in its generalized form is actually blind to the directional structure in the network: Q is left unchanged if we replace $\pi_x p_{xy}$ by its symmetrized version $\pi_x p^s_{xy} = \frac{1}{2}(\pi_x p_{xy} + \pi_y p_{yx})$, which is not sensitive to edge directions: $\pi_x p^s_{xy}$ can be large if for example only the edge (xy) exists, independent of the length of the shortest path from y back to x. In other words, optimizing Q can produce modules consisting of nodes that communicate only in one direction, which contradicts the intuition that all nodes in a module should be similar.

In this article, we require nodes in a module to communicate in both directions via short paths, and we will make this notion precise by introducing a novel measure

of communication I_{xy} between nodes based on cycles. Despite the fact that I_{xy} is symmetric, we will show that it inherits all necessary information about directions and structure of the original network. It will allow us to (a) propose a modified version $\bar{Q}$ of the modularity function which is sensitive to directional information and (b) transform G into an undirected network were x and y are connected by an edge with weight I_{xy} if $I_{xy} > 0$. On the transformed network we can then use any fuzzy clustering method designed for undirected networks.

Cycle decomposition and communication between nodes. We now briefly introduce the theory of cycle decompositions for Markov chains as developed in [18] and [19], which we will use for the RW process (1). An n-cycle (or n-loop) on G is defined as an ordered sequence [20] of n connected nodes $\gamma = (x_1, x_2, \ldots, x_n)$, whose length we denote by $|\gamma| = n$. Let $\mathcal{C}$ be the collection of simple (i.e. no self-intersections are allowed) cycles on G. Let $(X_n)_{1\leq n\leq T}$ be a realization of the Markov chain. We say that $(X_n)_{1\leq n\leq T}$ *passes through* the edge (xy) if $\exists n < T$ such that $X_n = x$ and $X_{n+1} = y$, and following [18, 19] we say that $(X_n)_{1\leq n\leq T}$ passes through γ if it passes through all edges of γ in the right order, but not necessarily consecutively. Let N_T^γ be the number of times γ is passed through up to time T. Then the limit

$$w(\gamma) := \lim_{T\to\infty} \frac{N_T^\gamma}{T} \tag{3}$$

exists almost surely [19] and gives us a uniquely defined probabilistic *cycle decomposition*, that is a collection $\Gamma = \{\gamma \in \mathcal{C} | w(\gamma) > 0\}$ of cycles with positive weights $w(\gamma)$ such that for every edge $(xy) \in E$ the flow decomposition formula holds:

$$F_{xy} = \sum_{\gamma \supset (xy)} w(\gamma) \tag{4}$$

where $F_{xy} = \pi_x p_{xy}$ is the flow through (xy) and we write $\gamma \supset (xy)$ if the edge (xy) is in γ. An explicit but impractical formula to calculate the weights $w(\gamma)$ was given in [19]. Alternatively, and in analogy to how one samples the transition matrix from a realization $(X_n)_{1\leq n\leq T}$, we can also sample them by obtaining the counts N_T^γ. In [19] such a sampling algorithm was described in detail.

Example: The barbell graph. As an example consider the barbell graph presented in Figure 1, with edge weight $K(l_0, r_0) = K(r_0, l_0) = \varepsilon < 1$, all other edge weights are one. Since every edge belongs to exactly one of the three loops $\alpha_c = (l_0, r_0)$, $\alpha_l = (l_0, l_1, \ldots, l_{n-1})$ and $\alpha_r = (r_0, r_1, \ldots, r_{n-1})$, the weights of these loops can be inferred directly from (4):

$$w(\alpha_l) = w(\alpha_r) = \frac{1}{2(n+\varepsilon)} =: w, \quad w(\alpha_c) = \varepsilon w.$$

We now define a measure for communication between nodes $x, y \in V$, as experienced by the realization $(X_n)_{1\leq n\leq T}$. To this end, think of $(X_n)_n$ as describing a 'postman' delivering parcels. Each time x is visited, a parcel is picked up, which is then delivered when y is reached. Let $N_T(x \to y)$ be the number of deliveries from x to y in time T - note that x and y need not be adjacent. Then

$$k_{xy} = \lim_{T\to\infty} \frac{N_T(x \to y)}{T} \tag{5}$$

is a measure for the communication between x and y, as experienced by $(X_n)_n$. This quantity is called *reaction rate* in Transition Path Theory [21]. It is easy to see that $k_{xy} = k_{yx}$, which does not mean that k_{xy} ignores directional information. In fact, k_{xy} contains information about all possible ways to go from x to y together with information about all possible ways to go from y to x, since the RW must return to x from y before going to y again.

Now focus on a single event $x \to y \to x$, which is realized by passing through a loop γ. This event gives $|\gamma|$ counts to $\sum_{y'} N_T(x \to y')$, one for each $y' \in \gamma$. In other words, the single event $x \to y \to x$ along γ is counted as one parcel picked up, but $|\gamma|$ parcels delivered. This 'overcounting' leads to $\sum_y k_{xy}$ not being meaningfully normalized.

To arrive at a normalized measure for communication between nodes, we will let the RW pick up the parcel at x and keep it until it returns to x, passing through a loop γ. Then it will be delivered to any node on γ with probability $\frac{1}{|\gamma|}$. Now the number $\tilde{N}_T(x \to y)$ of deliveries from x to y is

$$\tilde{N}_T(x \to y) = \sum_{\gamma \ni x,y} \frac{1}{|\gamma|} N_T^\gamma. \tag{6}$$

Finally, we define the *communication intensity* I_{xy} as the average rate of parcel deliveries:

$$I_{xy} := \lim_{T\to\infty} \frac{\tilde{N}_T(x \to y)}{T} = \sum_{\gamma \ni x,y} \frac{w(\gamma)}{|\gamma|}, \tag{7}$$

using (3) and (6), and we arrive at the probabilistic cycle decomposition introduced earlier. Intuitively, I_{xy} is large if there are many cycles connecting x and y, and if they are important ($w(\gamma)$ large) and short ($|\gamma|$ small). Thus, our new node communication intensity measure I_{xy} encodes the directional information. Additionally, I_{xy} is normalized as

$$\sum_y I_{xy} = \sum_y \sum_{\gamma \ni x,y} \frac{w(\gamma)}{|\gamma|} = \sum_{\gamma \ni x} w(\gamma) = \pi_x, \tag{8}$$

where the last equality follows from (4). This allows us to introduce the *loop transition matrix* $\bar{P}$ with components

$$\bar{p}_{xy} = \frac{I_{xy}}{\pi_x} = \frac{1}{\pi_x} \sum_{\gamma \ni x,y} \frac{w(\gamma)}{|\gamma|}. \tag{9}$$

Note that $\bar{P}$ is reversible since $I_{xy} = I_{yx}$ and it has the same stationary distribution as P, namely π. Using I_{xy} as a measure for the communication between nodes has provided us with a way to symmetrize P without loosing the directional information.

Generalized modularity function. We now define the modified modularity function

$$\bar{Q} = \sum_m \sum_{x,y} \chi_m(x) \left[\pi_x \bar{p}_{xy} - \pi_x \pi_y\right] \chi_m(y) \tag{10}$$

which encourages two nodes x,y to belong to the same module if I_{xy} is large, that is if x and y are connected by many short cycles with large weights $w(\gamma)$. The advantage of $\bar{Q}$ compared to Q is that it explicitly requires nodes in the same module to be connected in *both* directions via short paths.

However, optimizing $\bar{Q}$ will not result in fuzzy partitions. Instead we note that I_{xy} also gives us a transformation of G into an undirected, weighted network G_U, where we connect two nodes x, y by an edge with weight I_{xy} if $I_{xy} > 0$. Now we can use any fuzzy clustering algorithm for undirected networks on G_U. Notice that the standard transition matrix (1) on this network coincides with $\bar{P}$.

Fuzzy clustering algorithm for directed networks. We now present our new **LOop-based Lumping Algorithm (LOLA)** for fuzzy clustering of directed weighted networks:

(A) Given edge weights $K(x, y)$, construct P according to (1) and generate a realization $(X_n)_{1\le n\le T}$ of the Markov chain given by P.

(B) Use the sampling algorithm described in [19] to obtain the cycle decomposition Γ with weights $w(\gamma)$ via (3) and construct G_U using (7).

(C) Use any fuzzy clustering method on G_U.

In this paper, we will use the *Markov State Model (MSM) clustering* method [22, 23] in (C). MSM clustering identifies module cores C_m as the metastable sets of the RW process on G_U, which has $\bar{P}$ as it's transition matrix. Fuzzy affiliation functions are obtained as

$$q_m(x) = \mathbb{P}(X_t \text{ hits } C_m \text{ next}|X_0 = x), \quad \forall x \in V.$$

MSM clustering relies on the reversibility of $\bar{P}$ [24, 25] and can therefore not be applied to directed networks a priori.

Remark 1. If G is very large and/or very metastable, the sampling in step (B) can become too slow. To increase performance, we can replace the sampling algorithm in step (B) by any deterministic algorithm yielding a different cycle decomposition $(\Gamma, \tilde{w})$ which will still satisfy (4), but not (3), and will no longer be unique. In [18] a deterministic algorithm was given which can find a decomposition in polynomial time by iteratively reducing the flow F, and the matrix $\bar{P}(\Gamma)$ with components

$$\bar{p}_{xy}(\Gamma) = \frac{1}{\pi_x} \sum_{\gamma\in\Gamma,\, \gamma\ni x,y} \frac{\tilde{w}(\gamma)}{|\gamma|},$$

is still reversible with stationary distribution π.

Remark 2. If the network is undirected, every edge gives rise to a 2-loop. A cycle decomposition using only these loops exists by putting $\tilde{w}((xy)) = \pi_x p_{xy}$, and a short calculation shows $\bar{P}(\Gamma) = \frac{1}{2}(I + P)$. For *MSM clustering*, this means that the RW is only slowed down, but will produce the same modules.

We now illustrate the method by investigating two small example networks and a real-world network constructed from earthquake data.

The barbell graph cont'd. We can calculate I_{xy} explicitly:

$$I_{xy} = \begin{cases} w/n & x, y \in \alpha_l \\ w/n & x, y \in \alpha_r \\ \varepsilon w/2 & x \neq y \in \alpha_c \\ \varepsilon w/2 + w/n & x = y \in \alpha_c \end{cases}$$

Clustering this graph with LOLA produces a full partition into two modules $C_1 = \alpha_l$ and $C_2 = \alpha_r$, see Figure 1. We will call this partition the reference partition. The scores assigned by Q and $\bar{Q}$ to it are almost identical:

$$Q(C_1, C_2) = \frac{1}{2} - \frac{\varepsilon}{n+\varepsilon}, \quad \bar{Q}(C_1, C_2) = \frac{1}{2} - \frac{1}{2}\frac{\varepsilon}{n+\varepsilon}.$$

Now we consider what happens if one module is split in two chains of equal size $C_{1,a}$ and $C_{1,b}$. The total change ΔQ in Q resp. $\Delta\bar{Q}$ in $\bar{Q}$ under this split is

$$\Delta Q = \frac{1}{8} - \frac{1}{n+\varepsilon}, \quad \Delta\bar{Q} = \frac{1}{8} - \frac{1}{4}\frac{n}{n+\varepsilon}.$$

One can check that $\Delta Q > 0$ for $n \geq 8$ and $\Delta\bar{Q} < 0$ as long as $n > \varepsilon$. That is, as n grows Q favors a partition into more and more subchains with less then 8 nodes over the reference partition, even though increasing n actually increases the metastability of the reference partition. In contrast, the small chains favored by Q are not metastable at all. On the other hand $\bar{Q}$ always favors the reference partition.

Finally, one can check that if directions are ignored, Q behaves in exactly the same way as before, while $\bar{Q}$ favors a partitioning into subchains with less then 6 nodes since the nodes in these subchains can now communicate directly. Thus, in this example we can see that Q is not able to detect a qualitative difference between the undirected and directed barbell graph, while $\bar{Q}$ is.

False module identification. The next example is a network with 21 nodes, for which LOLA-clustering finds two metastable modules: C_1 colored in blue and C_2 colored in red, see Figure 2a). The rest of the network forms

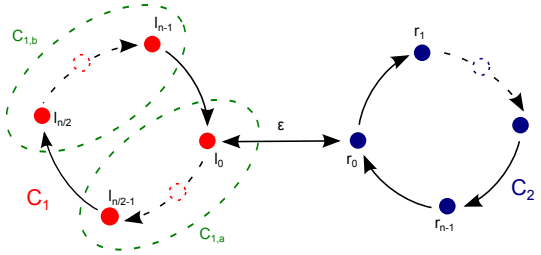
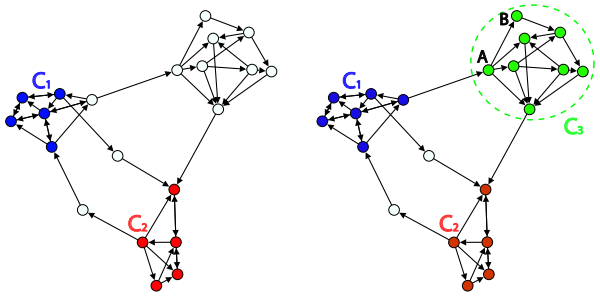


FIG. 1: The barbell graph, consisting of two loops with n nodes joined by an edge with weight ε.

a large transition region consisting of nodes with affiliation less than 0.8 (see Supplementary material for more detailed analysis). If we cluster this network using the gNG algorithm, a third module C_3 appears (green in 2b). However, C_3 is not a metastable module because none of its nodes are connected via short paths in both directions. For example, A and B are connected by a directed edge (AB), but in order to go from B to A, the RW has to pass through the whole network. Consequently, I_{AB} is small, and therefore LOLA-clustering overcomes this problem and is thus improving upon existing methods.



(a) (b)

FIG. 2: Example with false module identification. (a) Clustering produced by LOLA. (b) Clustering produced by gNG algorithm.

Earthquake graph. The last example is a timeseries $(X_n)_n$ of seismic events in California from 1952 to 2012, obtained from the SCEC[26]. A weighted, directed network is constructed from this data as follows: Only events with magnitude larger than $m_c = 2.5$ are considered (these are 48669 events). Space is discretized into boxes of $\Delta l = 0.1°$ in the latitude and longitude directions, and the boxes in which earthquakes occurred are the nodes. Then edge weights are defined as $K(x, y) = \mathbf{P}(X_n = x, X_{n+1} = y)$, which is estimated from the timeseries by counting successive events. See [27–30] for a discussion of this and related approaches to construct networks out of seismic data.

The presence of the timeseries data allows us to achieve a significant speed up by constructing the cycle decomposition in step (B) directly from the data. This entirely avoids the construction of P, and no further sampling is needed. For the network in question which has 2175 nodes and 28839 edges, step (B) takes 8 minutes on a standard laptop and reports 7739 cycles. This speedup will always be possible if the network is constructed from timeseries data, which is very common. We give more details in the supplementary material available at [url].

The fuzzy clustering obtained by LOLA is shown in Figure 3, where nodes are colored if they are assigned to a module with affiliation at least 0.8, and otherwise assigned to the transition region and shown grey. In fact 80% of the nodes are assigned to the transition region, but these correspond to only 25% of all events. This illustrates that our fuzzy clustering correctly reflects the uncertainty coming from limited data. A full clustering would have to cluster the grey nodes as well, even though not enough data is available to do so. LOLA-clustering finds 9 modules, all of which correspond to important faults or groups of faults, the largest one containing the San Andreas fault. Although the idea of using networks for analyzing seismic data is not new, to the best of our knowledge, our method is the first that can identify faults based only on the data and without using any additional information.

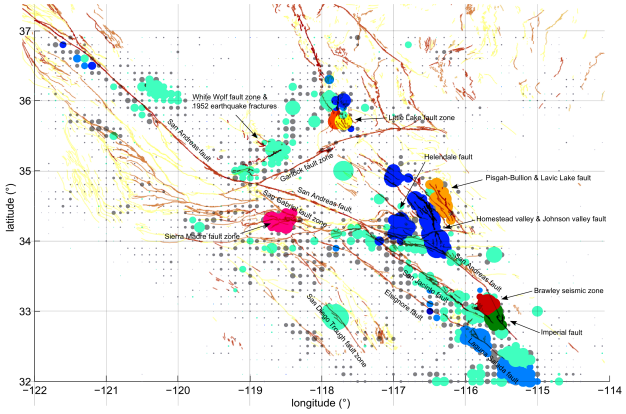


FIG. 3: Quaternary faults [31] in Southern California and the clustering of the SCEC timeseries found by LOLA. Node size is proportional to the number of events, color indicates the modules found.

Conclusion. In this paper we addressed the problem of module detection in weighted directed networks by constructing a novel measure of communication between nodes. This measure is based on a cycle decomposition of the probability flow which encodes the directional information. Since our measure is symmetric, it allows us to apply clustering methods designed for undirected graphs. We applied our new method on analyzing two toy examples, and a real-world directed earthquake network and showed how it overcomes the essential drawbacks of com-

mon methods.

We thank Stefan Rüdrich for valuable insights on earthquake data analysis and useful feedback on the manuscript, and Marco Sarich for helpful discussions.

* ralf.banisch@fu-berlin.de
† Christof.Schuette@fu-berlin.de
‡ djurdjev@fu-berlin.de
§ Zuse Institute Berlin

[1] M. Newman, A. Barabasi, and D. Watts, *The Structure and Dynamics of Networks* (Princeton Univ Press, Princeton, NJ, 2006).
[2] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[3] S. van Dongen, *Graph Clustering by Flow Simulation*, Ph.D. thesis, University of Utrecht (2000).
[4] P. Deuflhard and M. Weber, Linear Algebra and its Applications **398**, 161 (2005), special Issue on Matrices and Mathematical Biology.
[5] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
[6] M. E. J. Newman, SIAM Review **45**, 167 (2003).
[7] Y. Kim, S.-W. Son, and H. Jeong, Phys. Rev. E **81**, 016103 (2010).
[8] B. E. Latouche, P. and C. Ambroise, Statistical Modelling **12**, 93 (2012).
[9] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[10] J. M. Hofman and C. H. Wiggins, Phys. Rev. Lett. **100**, 258701 (2008).
[11] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[12] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[13] M. E. J. Newman and M. Girvan, Phys. Rev. E **69 (026113)** (2004).
[14] A. F. A Arenas1, J Duch and S. Gómez, New J. Phys. **9**, 176 (2007).
[15] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, ArXiv Physics e-prints (2006).
[16] M. E. J. Newman, Proceedings of the National Academy of Sciences **103**, 8577 (2006).
[17] A. Clauset, M. E. J. Newman, and C. Moore, Phys.Rev.E **70**, 066111 (2004).
[18] S. L. Kalpazidou, *Cycle Representations of Markov Processes* (Springer, 2006).
[19] D. Jiang, M. Qian, and M.-P. Quian, *Mathematical theory of nonequilibrium steady states: on the frontier of probability and dynamical systems.* (Springer, 2004).
[20] More precisely, cycles are equivalence classes of ordered sequences up to cyclic permutations. In this note we do not distinguish between cycles and their representatives.
[21] P. Metzner, C. Schütte, and E. Vanden-Eijnden, Multiscale Modeling & Simulation **7**, 1192 (2009).
[22] N. Djurdjevac, S. Bruckner, T. O. F. Conrad, and C. Schütte, Journal of Numerical Analysis, Industrial and Applied Mathematics **6**, 29 (2011).
[23] M. Sarich, N. Djurdjevac, S. Bruckner, T. O. F. Conrad, and C. Schütte, Journal of Computational Dynamics, In Press (2014).
[24] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, Comm. Math. Phys. **228**, 219 (2002).
[25] N. Djurdjevac, M. Sarich, and C. Schütte, Multiscale Modeling & Simulation **10**, 61 (2012).
[26] Southern California Earthquake Center, `www.scec.org`.
[27] S. Abe and N. Suzuki, Nonlinear Processes in Geophysics **13**, 145 (2006).
[28] S. Abe and N. Suzuki, EPL (Europhysics Letters) **65**, 581 (2004).
[29] J. Davidsen, P. Grassberger, and M. Paczuski, Phys. Rev. E **77**, 066104 (2008).
[30] K. F. Tiampo, J. B. Rundle, W. Klein, J. Holliday, J. S. Sá Martins, and C. D. Ferguson, Phys. Rev. E **75**, 066107 (2007).
[31] *US Geological Survey* (http://www.earthquake.usgs.gov).