Publication V

János Török, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo K. Kaski, and János Kertész. Opinions, Conflicts and Consensus: Modeling Social Dynamics in a Collaborative Environment. *Physical Review Letters*, Volume 110, Issue 8, 088701, February 2013.

Opinions, Conflicts, and Consensus: Modeling Social Dynamics in a Collaborative Environment

János Török,[1] Gerardo Iñiguez,[2] Taha Yasseri,[3,1] Maxi San Miguel,[4] Kimmo Kaski,[2] and János Kertész[5,1,2]
[1] *Institute of Physics, Budapest University of Technology and Economics, H-1111 Budapest, Hungary*
[2] *Department of Biomedical Engineering and Computational Science, FI-00076 Aalto, Finland*
[3] *Oxford Internet Institute, University of Oxford, OX1 3JS Oxford, United Kingdom*
[4] *IFISC (CSIC-UIB), Campus Universitat Illes Balears, E-07071 Palma de Mallorca, Spain*
[5] *Center for Network Science, Central European University, H-1051 Budapest, Hungary*

(Received 19 July 2012; published 19 February 2013)

Information-communication technology promotes collaborative environments like Wikipedia where, however, controversy and conflicts can appear. To describe the rise, persistence, and resolution of such conflicts, we devise an extended opinion dynamics model where agents with different opinions perform a single task to make a consensual product. As a function of the convergence parameter describing the influence of the product on the agents, the model shows spontaneous symmetry breaking of the final consensus opinion represented by the medium. In the case when agents are replaced with new ones at a certain rate, a transition from mainly consensus to a perpetual conflict occurs, which is in qualitative agreement with the scenarios observed in Wikipedia.

DOI: 10.1103/PhysRevLett.110.088701 PACS numbers: 89.65.Ef, 05.65.+b, 64.60.an, 89.75.Fb

Society represents a paradigmatic example of complex systems, where interactions between many constituents and feedback and other nonlinear mechanisms result in emergent collective phenomena. The recent availability of large data sets due to information-communication technology has enabled us to apply more quantitative methods in social sciences than before. Physicists play an increasing role in the study of social phenomena by applying physics concepts and tools to investigate them [1].

Social interactions are heavily influenced by the opinions of the members of the society. This is especially true when complex tasks are to be solved by cooperation, as practiced throughout the history of mankind. In this respect, new technologies open up unprecedented opportunities: by using the Internet and related facilities, even remote members of large groups can work on the same task and achieve a higher level of synergy. Examples include open software projects or large collaborative scientific endeavors like high-energy physics experiments. However, it is unavoidable that in such cases differences in attitudes, approaches, and emphases (in short, opinions) occur. Then questions arise: How can a task be solved in a collaborative environment of agents having diverse opinions? How do conflicts emerge and get resolved? The understanding of these mechanisms may lead to an increase in efficiency of value production in cooperative environments. A prime example of the latter is Wikipedia, a free, Web-based encyclopedia project where volunteering individuals collaboratively write and edit articles on their desired topics. Wikipedia is particularly well-suited also as a target for a wide range of studies, since all changes and discussions are recorded and made publicly available [2–7].

Recently, the controversy and dynamical evolution of Wikipedia articles have been studied in detail and typical patterns of different categories of the so-called *edit wars* have been identified [8–12]. Take for example Fig. 1, where the evolution of a controversy measure M based on mutual reverts and maturity of editors is shown for three different regimes of conflict.

Our aim in this Letter is to model controversy and conflict resolution in a collaborative environment and, where possible, to qualitatively compare our results with different scenarios observed in the Wikipedia. One of the most developed areas of quantitative modeling in social phenomena is opinion dynamics [1,13,14]. These models have much in common with those of statistical physics, yet interactions are socially motivated. The problem with these models is, similarly to evolutionary game theoretic ones [15], that results are usually evaluated by qualitative subjective judgment instead of comparison with empirical data, one exception being the study of elections [16,17]. The well-documented Wikipedia edit wars can also be of use in this respect.

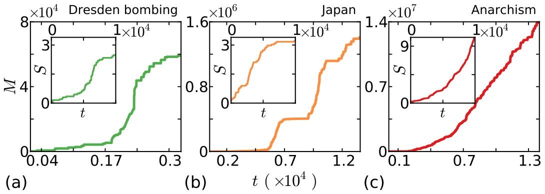

FIG. 1 (color online). Empirical controversy measure M [10] as a function of the number of edits t for three different conflict scenarios in Wikipedia, corresponding to: (a) single conflict, (b) plateaus of consensus, and (c) uninterrupted controversy. Titles are the article topics. Inset: Theoretical conflict measure $S(t)$ of Eq. (2).

Our model is based on the *bounded confidence* (BC) mechanism [18,19], describing a generic case when convergence occurs only upon opinion difference being smaller than a given threshold. Here, N agents are characterized by continuous opinion variables $x_i \in [0, 1]$ and their interactions are pairwise. The agents' mutual influence is controlled by the convergence parameter μ_T, only if their opinions differ less than a given tolerance ϵ_T; i.e., for $|x_i - x_j| < \epsilon_T$, we update as follows,

$$(x_i, x_j) \mapsto (x_i + \mu_T[x_j - x_i], x_j + \mu_T[x_i - x_j]). \qquad (1)$$

The dynamics set by Eq. (1) has been studied extensively in the literature [1], initially by using the mean-field approach of two-body inelastic collisions in granular gases [20,21]. It leads to a frozen steady state which is characterized by $n_c \sim 1/(2\epsilon_T)$ disjoint opinion groups. This is caused by the instability in the initial opinion distribution near the boundaries. Also, n_c increases as $\epsilon_T \rightarrow 0$ in a series of bifurcations [22]. The BC mechanism has many extensions, such as vectorial opinions [23] and coupling with a constant external field [24].

Let us now consider the case in which agents with different opinions have the task to form a consensual product. For this, we can couple BC with a medium considered as the common product on which the agents should work collectively. In this case, the medium has also a convergence parameter $\mu_A \in [0, 1]$ and tolerance $\epsilon_A \in [0, 1]$, such that the opinion $A \in [0, 1]$ represented by the medium can be modified by agents being dissatisfied with it. If $|x_i - A| > \epsilon_A$, we update $A \mapsto A + \mu_A(x_i - A)$. Conversely, agents tolerating the current state of the common product ($|x_i - A| \leq \epsilon_A$) adapt their view towards the medium, $x_i \mapsto x_i + \mu_A(A - x_i)$. Thus, agents can interact directly with each other and indirectly through the medium, and a complex dynamics governed by competition of these local (direct) and global (indirect) interactions emerges. Such an interplay is also present in many other systems, ranging from surface chemical reactions [25] and sand dunes [26] to arrays of chaotic electrochemical cells [27].

All opinions are, first, initialized uniformly at random and the original BC algorithm is run until opinion groups are formed. Then, N pairs of agent-agent and agent-medium interactions are performed in each time step t, with agents being selected uniformly at random. If all agents fall within the tolerance level of the medium, the dynamics is frozen and we call such stable state *consensus*. The cumulative amount of conflict or controversy in the system is defined as the total sum of changes in the medium,

$$S(t) = \sum_{t'=1}^{t} \sum_{i=1}^{N} |A(i) - A(i-1)|. \qquad (2)$$

This quantity is analogous to the empirical controversy measure M as it sums up the actions of dissatisfied agents [10].

In what follows, we will analyze two versions of this model: (i) with the agent pool fixed and (ii) with agents being replaced by new ones at a certain rate. For the sake of simplicity, we fix $\epsilon_T = 0.2$ (leading in general to one large mainstream and two small extremist groups) and $\mu_T = 0.5$ (implying a fair compromise of opinions).

Fixed agent pool.—For finite N and if $0 < \epsilon_A, \mu_A < 1$, the system always reaches consensus. Let i be the agent with the largest opinion x_i, so that a discussion with any other agent may only lower the value of x_i. Consider now, the event in which agent i alone modifies the medium for a number of consecutive steps. If $A + \epsilon_A < x_i$, the medium is moved towards the opinion of the agent by a finite amount $\mu_A(x_i - A)$. Finally, after a finite number of steps when x_i falls within the tolerance level of the medium, x_i will be lowered by a finite amount larger than $\mu_A\epsilon_A(1 - \mu_A)$. In this way, we have devised an event of finite probability where x_i is decreased, which in turn, leads to a shrunken interval for the available opinion pool. Thus, the convergence to consensus is secured and the relaxation time τ can be defined, which, however, may be astronomical for large N.

If the tolerance ϵ_A is large, however, consensus may be quickly reached in a finite number of unidirectional steps. By decreasing ϵ_A, a limiting case is reached, where the opinion of the medium starts to oscillate between the points $1 - \epsilon_A$ and ϵ_A. The change in the opinion of the medium should be the distance between such points, so for a given μ_A, the limit of oscillatory behavior is $\epsilon_A^* \equiv 1/(2 - \mu_A)$. From now on, we are interested in the nontrivial case $\epsilon_A < \epsilon_A^*$.

We observed three different scenarios (see Fig. 2) for the dynamics depending on the values of ϵ_A and μ_A: In case I, the opinion of the mainstream group fluctuates for a long time around a stable value. This state is characterized by an astronomical relaxation time. In case II, the opinion of the mainstream group oscillates between the vicinities of the extremists, with τ independent of N. In case III, the extremists converge as groups towards the mainstream opinion.

The transition between regimes I and II can be described with stability analysis. We use the following assumptions:

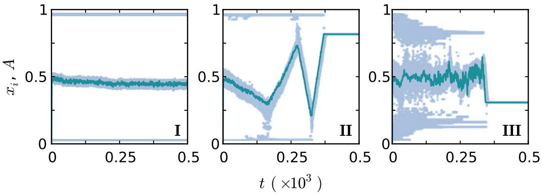

FIG. 2 (color online). Example evolutions of the agents' opinions x_i (light gray or blue) and the value A of the medium (dark gray or green) for three qualitatively different regimes, corresponding to $(\epsilon_A, \mu_A) = (0.075, 0.2)$ in I, (0.075, 0.45) in II, and (0.15,0.7) in III.

First, there are three opinion groups, one mainstream with opinion x_0 and two extremists with opinions x_- and x_+. Second, N is large enough such that the change in the opinions of the groups and correspondingly, the change in the probability distribution ρ_A of the opinion of the medium is slow compared to a single edit. In this case, the stationary master equation for ρ_A can be written as

$$0 = \sum_i (-\rho_A(A)\Theta[d_{A,x_i}] + \rho_A(A_i)\Theta[d_{A_i,x_i}])n_i, \quad (3)$$

where n_i is the relative size of group $i \in \{0, -, +\}$, $d_{A,x_i} = |A - x_i| - \epsilon_A$, and $A_i = (A - \mu_A x_i)/(1 - \mu_A)$ is the opinion of the medium from where it would jump to A after the interaction with x_i. For all values of A with $|x_0 - A| < \epsilon_A$, the mainstream group is moved towards A with probability $n_0\rho_A$. The resulting velocity of the opinion of the mainstream group is then

$$v_0(x_0) = V(\mu_A)n_0 \int_{x_0-\epsilon_A}^{x_0+\epsilon_A} \rho_A(A)(A - x_0)dA, \quad (4)$$

where $V(\mu_A)$ is a positive constant. In Fig. 3(a), we show how v_0 depends on μ_A. If $n_- = n_+$, the opinion of the mainstream group is stable at $x_0 = 1/2$ for low values of μ_A, due to the negative slope of v_0. Its point of stability bifurcates as μ_A increases and the mainstream group will drift towards one of the extremes. As soon as the opinions of the extremists get within the tolerance of the medium, some of them will move towards the mainstream. When enough extremists have converted to the mainstream group, the velocity of the mainstream gets reverted [see dashed line in Fig. 3(a)] and the mainstream group will head towards the other extreme. According to our calculations, more than 25% population difference between extremists is needed for the reversal. This ensures that consensus is reached after a few oscillations, which makes the relaxation time independent of N. In Fig. 3(b), the numerical boundary between regimes I and II is drawn at the marginal stability of the mainstream group. We note here, that for some values of ϵ_T, the shape of the boundary between regime I and II is more complicated and may even include islands.

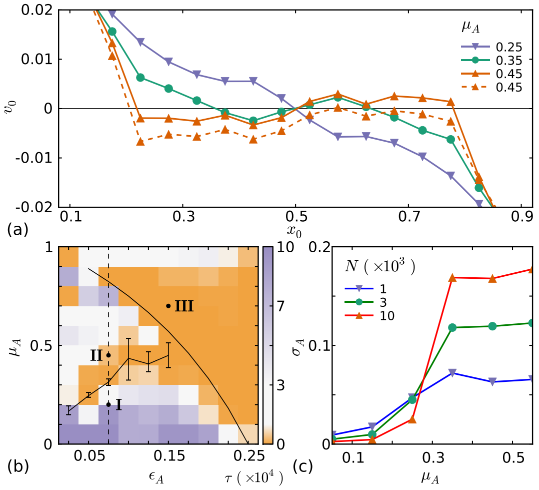

FIG. 3 (color online). (a) Velocity v_0 of the mainstream group at $\epsilon_A = 0.075$ as a function of its opinion x_0, for equal extremist group sizes (solid lines) and for 25% more extremists at x_- (dashed line). (b) Phase diagram (ϵ_A, μ_A) with shading indicating the relaxation time τ. Points give parameter values for the examples in Fig. 2 and lines indicate boundaries between different regimes, denoted by Roman numerals. (c) Order parameter σ_A for the transition between I and II at $\epsilon_A = 0.075$ [dashed line in (b)].

After the last interaction with the extremists, the opinion of the medium and the mainstream group will remain in the vicinity of one of the extremes. In the thermodynamic limit, this leads to a symmetry breaking in the stationary state of regime II. Conversely, in regime I the relaxation time grows exponentially with the number of agents, as a sequence of low probability ($\propto 1/N$) events are needed for convergence. Thus, for $N \rightarrow \infty$, we find a stationary state where the opinion of the mainstream group is at 1/2. Small shifts are possible due to differences in extremist populations, but since their ratio determines the opinion of A, disturbances vanish as $1/\sqrt{N}$. Then, it is safe to define the order parameter σ_A as the standard deviation of the opinion of the mainstream group. When $N \rightarrow \infty$, this tends to 0 for case I and increases for case II, as depicted in Fig. 3(c). The latter reflects a bimodal distribution ρ_A corresponding to the broken symmetry.

Regime III is characterized by converging extremist groups. As ϵ_A and μ_A increase, the jump of the medium is big enough so that in one step, the extremists get within its tolerance interval and start drifting inwards. Thus, the step size must be $\Delta = \mu_A(1/2 - \epsilon_A) = 1/2 - 2\epsilon_A$, where 1/2 is the distance between the mainstream and the extremist groups. The boundary $\mu_A = 1 - \epsilon_A/(1/2 - \epsilon_A)$ is shown in Fig. 3(b), separating regime III from the rest.

Agent replacement.—In real systems, the agent pool is often not fixed in time as people come and go. We introduce the agent renewal rate p_{new} as the probability for an agent to be replaced by a new one with random opinion before the interactions. In this section, we fix $\mu_A = 0.1$ to reduce the number of parameters, focusing solely on N, ϵ_A, and p_{new}. Intuitively, it is then clear that for $\epsilon_A < 1/2$ and $p_{\text{new}} > 0$, the dynamics never converges to a stationary state, as opposed to the case of a fixed agent pool. This is because for any A value, there is a finite probability that a new agent enters with an opinion outside the tolerance level of the medium, after which this new agent may change the value of A.

In the insets of Fig. 1, we display some examples of the time evolution of S for $Np_{\text{new}} = 4$ and $\epsilon_A = 0.47, 0.46, 0.44$.

As in Ref. [10], we can distinguish three qualitatively different regimes as ϵ_A decreases: (a) Single conflict, where S is dominated by an initial increase and a prolonged peace signaled by long plateaus. (b) A series of small plateaus of consensus separated by conflicts. (c) A continuous increase of S indicating a permanent state of war. We define a conflict as the period between two plateaus of S and denote the number of conflicts per unit time by r. The two extreme regimes are both characterized by low values of r: in the peaceful regime, where ϵ_A is large, there are few conflicts, while for small ϵ_A, there is only one never-ending conflict. These regimes are separated by a region full of small conflicts.

In the inset of Fig. 4, we show the variation of r with ϵ_A. As N increases, regimes (a) and (c) are indeed separated by a thinning transition region (b) of many conflicts. A critical tolerance value between consensus and controversy regimes is then identified by a maximum in the conflict density r and is denoted by $\bar{\epsilon}_A$.

We note that both $\bar{\epsilon}_A$ and $r(\bar{\epsilon}_A)$ increase for larger N, but true divergence associated with critical behavior near a phase transition cannot be observed here due to the condition $r \leq 1$. The transition point $\bar{\epsilon}_A$ depends both on N and p_{new}, and its corresponding phase diagram is shown in Fig. 4. The transition between peace and conflict can be derived from the matching of two time scales: (i) the relaxation time of the system without agent renewal and (ii) the time scale of agent renewal. If the latter is too small, no relaxation takes place and we have an ever-present conflict. Thus, at the transition point, both time scales should be equal.

The relaxation time τ for a fixed agent pool can be calculated in regime III if $\epsilon_A > 0.25$ (for details, see Supplemental Material [28]). In this case, A can make only few jumps up and down. Knowing the distribution of the opinion of the agents the task is to eliminate the extremists. The rate equations for the medium and extremist movement can be established and solved analytically to give the mean relaxation time as

$$\tau = cN([2e^2 + e_0^2(n-1)]n - ee_0(n-1)(2+n)), \quad (5)$$

where $e = \epsilon_A^* - \epsilon_A$, $e_0 = \epsilon_A^* - 1/2$, c is a constant depending on μ_A, and n denotes the integer part of e/e_0 counting the number of steps the medium can make in one direction.

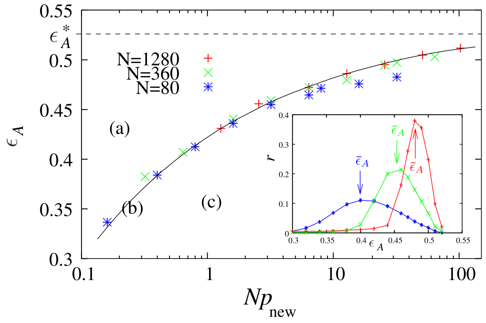

FIG. 4 (color online). Phase diagram (Np_{new}, ϵ_A) in the case of agent replacement. The transition point $\bar{\epsilon}_A$ shows agreement between numerical (points) and analytical (line) results. Letters denote regimes of conflict and correspond to labels in Fig. 1. Inset: Conflict rate r as a function of ϵ_A for varying N at $p_{\text{new}} = 0.01$.

The number of new agents per unit time is Np_{new}, so we expect the transition point $\bar{\epsilon}_A$ to be at $1 = Np_{\text{new}}\tau(\bar{\epsilon}_A)$, shown in Fig. 4 as a continuous line. Such a result is in considerable agreement with the numerical computation of $\bar{\epsilon}_A$, with one single fit parameter. It also holds for other values of μ_A, deviations appearing only for $\mu_A \ll 0.01$. We expect that in the $p_{\text{new}} \to 0$ limit $\bar{\epsilon}_A \equiv 0$, as there is no state with permanent conflict in the fixed agent pool case. Furthermore, for $p_{\text{new}} \to \infty$, we have $\bar{\epsilon}_A = \epsilon_A^*$, as this is the point above which no position of the medium allows for a conflict. The curves for different system sizes fall upon each other if the number of new agents per unit time is used as control parameter, irrespective of the total number of agents. This means that consensus is as vulnerable to many people as it is to few.

Overall, agent replacement for the nontrivial case $\epsilon_A < \epsilon_A^*$ gives a transition between regime (a) representing practically peace ($dS/dt \ll 1$) and regime (c) representing continuous conflict ($dS/dt \simeq 1$). The transition can happen if many new agents enter the system either by increasing p_{new} or N. An analogue of this transition is indeed observed in Wikipedia, namely that a peaceful article can suddenly become controversial when more people get involved in its editing [6,10].

Summary.—A general question addressed by our extended opinion formation model is the competition and feedback loop between direct agent-agent interactions and the indirect interaction of agents with a "mean field" collectively created. We find that convergence is always reached when the indirect interaction mechanism is present, even in situations in which the agent-agent interaction alone does not lead to it. We have also described different dynamical regimes of approaching convergence, finding a symmetry-breaking mechanism for the collectively created opinion A at a critical value of the convergence parameter μ_A. In the case of agent replacement, we find a transition from a relatively peaceful situation to a perpetual state of conflict when the rate of replacement is increased above a threshold. Such finding is in agreement with different conflict scenarios observed in Wikipedia, allowing us to translate some of its policies to hinder conflict into model parameters. Freezing editorial activity temporarily or banning inexperienced fighting editors (decrease Np_{new}), and moving disputed topics to "controversy" sections (increase ϵ_A), all lead to more peaceful articles. This first step in comparing an opinion model with real data calls to extend the model by including networked interactions between agents [29], individual activities, and tolerances with wide distributions, leader-follower dynamics [30], heterogeneously-distributed times between successive edits [31], and external events to enable a more quantitative comparison between the model and empirical observations.

J. T, T. Y, K. K, and J. K acknowledge support from EU's FP7 FET Open STREP Project ICTeCollective No. 238597, and J. K also support by FiDiPro program of TEKES. M. S. M acknowledges support from project FISICOS (FIS2007-60327) of MINECO. G. I. acknowledges support from an STSM within COST Action MP0801 and is grateful to IFISC for hospitality.

[1] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[2] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet, Phys. Rev. E **74**, 016115 (2006).
[3] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).
[4] D. M. Wilkinson and B. A. Huberman, First Monday **12** (2007).
[5] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07* (ACM, New York, 2007), p. 453.
[6] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, Phys. Rev. Lett. **105**, 158701 (2010).
[7] T. Yasseri, R. Sumi, and J. Kertész, PLoS ONE **7**, e30091 (2012).
[8] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész, in *Proceedings of the ACM WebSci'11* (ACM, Koblenz, 2011), p. 1.
[9] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész, in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, and 2011 IEEE Third International Conference on Social Computing* (IEEE, Boston, 2011), p. 724.
[10] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész, PLoS ONE **7**, e38869 (2012).
[11] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. Chang, in *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM '08* (ACM, New York, 2008), p. 171.
[12] U. Brandes and J. Lerner, Inform. Visual. **7**, 34 (2008).
[13] J. A. Hołyst, K. Kacperski, and F. Schweitzer, in *Annual Reviews of Computational Physics IX* (World Scientific, Singapore, 2001), Chap. 5, p. 253.
[14] H. Xia, H. Wang, and Z. Xuan, Int. J. Knowl. Syst. Sci. **2**, 72 (2011).
[15] G. Szabó and G. Fáth, Phys. Rep. **446**, 97 (2007).
[16] A. T. Bernardes, D. Stauffer, and J. Kertész, Eur. Phys. J. B **25**, 123 (2002).
[17] S. Fortunato and C. Castellano, Phys. Rev. Lett. **99**, 138701 (2007).
[18] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Adv. Compl. Syst. **03**, 87 (2000).
[19] R. Hegselmann and U. Krause, J. Artif. Soc. Soc. Simulat. **5**, 2 (2002).
[20] E. Ben-Naim and P. L. Krapivsky, Phys. Rev. E **61**, R5 (2000).
[21] A. Baldassarri, U. M. B. Marconi, and A. Puglisi, Europhys. Lett. **58**, 14 (2002).
[22] E. Ben-Naim, P. L. Krapivsky, and S. Redner, Physica (Amsterdam) **183D**, 190 (2003).
[23] S. Fortunato, V. Latora, A. Pluchino, and A. Rapisarda, Int. J. Mod. Phys. C **16**, 1535 (2005).
[24] J. C. González-Avella, M. G. Cosenza, V. M. Eguíluz, and M. S. Miguel, New J. Phys. **12**, 013010 (2010).
[25] G. Veser, F. Mertens, A. S. Mikhailov, and R. Imbihl, Phys. Rev. Lett. **71**, 935 (1993).
[26] H. Nishimori and N. Ouchi, Phys. Rev. Lett. **71**, 197 (1993).
[27] I. Z. Kiss, Y. Zhai, and J. L. Hudson, Phys. Rev. Lett. **88**, 238301 (2002).
[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.088701 for the derivation of the relaxation time.
[29] S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs, J. Inf. Tech. **25**, 230 (2010).
[30] T. Yasseri and J. Kertész, J. Stat. Phys. (to be published).
[31] J. Fernández-Gracia, V. M. Eguíluz, and M. S. Miguel, Phys. Rev. E **84**, 015103(R) (2011).