A

Activity Dynamics in Collaboration Networks

SIMON WALK*, Institute for Information Systems and Computer Media, Graz University of Technology
DENIS HELIC, Knowledge Technologies Institute, Graz University of Technology
FLORIAN GEIGL, Knowledge Technologies Institute, Graz University of Technology
MARKUS STROHMAIER, GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau

Abstract Many online collaboration networks struggle to gain user activity and become self-sustaining due to the ramp-up problem or dwindling activity within the system. Prominent examples include online encyclopedias such as (Semantic) MediaWikis, Question and Answering portals such as StackOverflow, and many others. Only a small fraction of these systems manage to reach self-sustaining activity, a level of activity that prevents the system from reverting to a non-active state. In this paper, we model and analyze activity dynamics in synthetic and empirical collaboration networks. Our approach is based on two opposing and well-studied principles: (i) without incentives, users tend to lose interest to contribute and thus, systems become inactive, and (ii) people are susceptible to actions taken by their peers (social or peer influence). With the activity dynamics model that we introduce in this paper we can represent typical situations of such collaboration networks. For example, activity in a collaborative network, without external impulses or investments, will vanish over time, eventually rendering the system inactive. However, by appropriately manipulating the activity dynamics and/or the underlying collaboration networks, we can jump-start a previously inactive system and advance it towards an active state. To be able to do so, we first describe our model and its underlying mechanisms. We then provide illustrative examples of empirical datasets and characterize the barrier that has to be breached by a system before it can become self-sustaining in terms of critical mass and activity dynamics. Additionally, we expand on this empirical illustration and introduce a new metric p—the *Activity Momentum*—to assess the activity robustness of collaboration networks.

Additional Key Words and Phrases: Dynamical Systems, Activity Dynamics, Network Science, Critical Mass, Activity Momentum, Collaboration Networks

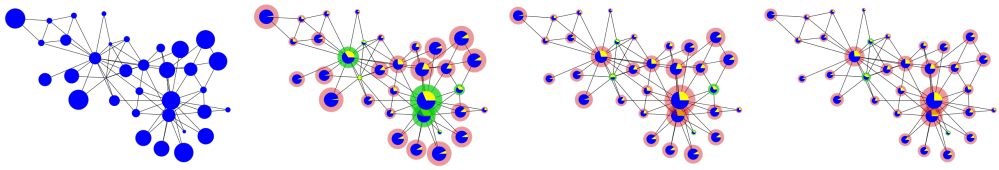
ACM Reference Format:
Simon Walk, Denis Helic, Florian Geigl and Markus Strohmaier, 2015. Activity Dynamics in Collaboration Networks. *ACM* V, N, Article A (January YYYY), 33 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

One of the major problems faced by both, new and existing online social and collaboration networks—such as Facebook or StackOverflow—revolves around efficiently identifying and motivating the appropriate users to contribute new content. In an optimal scenario, this newly contributed content provides enough incentive for other users to contribute, triggering further actions and contributions. Once such a self-reinforced state of increasing activity is reached, we can say that a system becomes self-sustaining, meaning that sufficiently high levels of activity are reached, which will keep the system active without further external impulses. For example, when looking at well-established collaborative websites, such as StackOverflow or Wikipedia, we already know that at some point in time, these systems have become self-sustaining (in terms of activity), evident in their steady growing number of supporters and overall activity.

However, these self-sustaining states are neither easy to reach nor guaranteed to last. For example, Suh et al. [2009] showed that the growth of Wikipedia is slowing down, indicating a loss in momentum and perhaps even first evidence of a collapse. Moreover, we typically lack the tools to properly analyze these trends in activity dynamics and thus, can not even

* Corresponding Author: S. Walk, Graz University of Technology, Institute for Information Systems and Computer Media, Inffeldgasse 16c/I, A-8010 Graz. E-Mail: simon.walk@tugraz.at; T: +43 (316) 873 - 5619.



(a) Intrinsic Activity (blue) and Peer Influence (yellow) at time t_0 (b) Intrinsic Activity (blue) and Peer Influence (yellow) at time t_1 (c) Intrinsic Activity (blue) and Peer Influence (yellow) at time t_2 (d) Intrinsic Activity (blue) and Peer Influence (yellow) at time t_3

Fig. 1: **Intrinsic Activity and Positive Peer Influence.** Activity dynamics in collaboration networks, represented by users as nodes, collaboration as edges and activity as node size (Figure (a)), are based on two opposing principles. The *Activity Decay Rate* postulates the loss of intrinsic activity (blue color of nodes) per user over time. In contrast, the *Peer Influence Growth Rate* follows the intuition, that users in collaboration networks are (positively) influenced by their peers (yellow color of nodes) where more active peers exercise a higher influence than less active peers. We initialize the network at time t_0 with random intrinsic activities. Nodes with a green halo at times t_1 to t_3 represent users that exhibit a gain in their overall activity between two iterations t_n and t_{n+1}, as the exercised positive peer influence is higher than the intrinsic loss of activity. Analogously, red halos represent decreases in overall activity. At first, very central (high degree) nodes with smaller activity values manage to increase their overall activity, while very active central nodes already start to lose activity. After t_3 or more iterations, due to overall decreasing activities and hence, decreasing peer influences, all nodes in the collaboration network eventually start to lose activity and inevitably converge towards zero activity.

perform such simple tasks as detecting self-sustaining system states. Therefore, we argue that new tools and techniques are needed to model, monitor and simulate activity dynamics for collaboration networks.

The high-level contributions of this work are two-fold. First, we introduce a model that is capable of simulating activity dynamics for online collaboration networks. Second, we describe in detail how to fit the model to empirical datasets, simulate trends in activity dynamics and interpret our findings. The proposed model is based on the formalism of continuous deterministic dynamical systems—meaning that activity is modeled by a system of coupled non-linear differential equations. Each user of the system is represented by a single quantity (the current activity), and the social ties between users define the coupling of variables. In general, when using dynamical systems on networks, we define the (micro-)behavior of each user to observe and gather new insights into the (macro-)behavior of the system. For a more detailed introduction to dynamical systems see Section 5 and Newman [2010]. For simplicity, we do not take individual differences between users into account—the dynamics and its parameters are the same for each user in the population. This allows us to configure the model with a single parameter, which is a ratio of the following two parameters, representing two basic activity mechanisms (cf. Figure 1) in online collaboration networks:

(i) *Activity Decay Rate* λ, which postulates how fast users lose interest to contribute,
(ii) *Peer Influence Growth Rate* μ, postulating to what extent users are influenced by the actions taken by their peers.

A first analysis of the model shows that activity dynamics in collaboration networks have an obvious and natural fixed point—the point of complete inactivity—where all contribu-

tions of the users have seized. However, by slightly manipulating the parameters in our model we show that it is possible to destabilize the fixed point, resulting in a potential increase of activity. We then outline the process of calculating the *Activity Decay Rate* and *Peer Influence Growth Rate* for existing collaboration networks, simulate their corresponding activity dynamics and expand our understanding of critical mass—via the notion of *System Mass* and *Activity Momentum*—in collaboration networks by interpreting our findings.

The remainder of this paper is structured as follows: In Section 2 we introduce and examine our model analytically. We then continue with the model illustration by simulating activity dynamics for a synthetic dataset and discuss different evolution scenarios of our parameters and their implications. In Section 3 we outline the process of applying our model on empirical datasets. In Section 4 we introduce the notion of *System Mass* and *Activity Momentum*, review related work in Section 5 and summarize our findings and discuss limitations and implications for future work in Section 6.

2. MODELING ACTIVITY DYNAMICS

We model activity dynamics in an online collaboration network as a dynamical system on a network. Hereby, the nodes of a network represent users of the system and links represent the fact that the users have collaborated in the past. We represent the network with an $n \times n$ adjacency matrix $\boldsymbol{A}$, where n is the number of nodes (users) in the network. We get $A_{ij} = 1$ if nodes i and j are connected by a link and $A_{ij} = 0$ otherwise. Since collaboration links are undirected, the matrix $\boldsymbol{A}$ is symmetric, thus $A_{ij} = A_{ji}$, for all i and j. We denote the total number of links in the network with m, and thus we have $2m = \sum_{ij} A_{ij}$.

We model activity as a continuous real-valued variable a_i evolving on node i of the network in continuous time t. The general time evolution equation can be written as follows (see also Newman [2010]):

$$\frac{da_i}{dt} = \underbrace{f_i(a_i)}_{\substack{\text{Intrinsic Activity} \\ \text{Evolution of i}}} + \overbrace{\sum_j A_{ij} \underbrace{g_i(a_i, a_j)}_{\text{Influence of j on i}}}^{\text{Peer Influence}}, \tag{1}$$

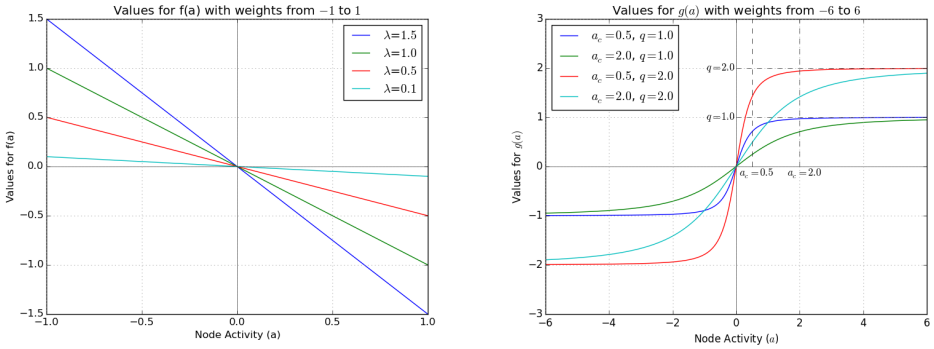
where $f(a_i)$ specifies the intrinsic activity evolution of node i and $g(a_i, a_j)$ describes the influence of neighbor j on node i. To simplify, we assume that the intrinsic activity dynamics as well as the influence of node neighbors are the same for each node i and for each neighbor pair (i, j). This means that we have a single intrinsic activity function $f(a_i)$ for all nodes i, as well as a single peer influence function $g(a_i, a_j)$ for all node pairs (i, j).

In addition, we make the following assumptions:

Intrinsic Activity Decay. Without external incentives or without positive influence from their social connections, each user has a tendency to slowly reduce activity. For example, people slowly lose interest to participate in collaborative networks or exhaust their resources. An observation that specifically reflects this inherent exhaust of activity over time has been made by Danescu-Niculescu-Mizil et al. [2013] for different online communities. We model this situation by using a linear function for $f(a_i)$:

$$f(a_i) = -\lambda a_i, \lambda > 0 \tag{2}$$

We call parameter λ the *Activity Decay Rate*—the rate at which users reduce their activity per unit time, given a complete absence of other (positive) incentives. The specific form of $f(a_i)$ results in an exponential decay ($a_i(t) = a_i(t_0)e^{-\lambda t}$, with $a_i(t_0)$ being the initial activity of node i at time t_0) of activity without any external influence. Thus, without other positive impulses the activity of every user will decay over time (see Figure 2(a)).



(a) Intrinsic Activity Decay (b) Extrinsic Peer Influence

Fig. 2: **Intrinsic Activity Decay** is the rate at which users reduce their activity per unit time and is represented as a linear function in the form of $f(a) = -\lambda a$, which results in an exponential decay in activity that converges towards zero. **Extrinsic Positive Peer Influence** describes to what extent users are influenced by the actions taken by their peers, and is represented as a monotonically increasing function of a users activity in the form of $g(a) = (qa)/\sqrt{a_c^2 + a^2}$. It naturally saturates at *Maximum Peer Activity Flow* q as activity reaches infinity and, in our simulations, can never be negative per definition (see Equation 3). When the user activity passes the point of the *Critical Activity Threshold* a_c, peer influence gains notable weight and influences neighbors to "do something" (become active).

Positive Peer Influence. People tend to copy their friends [Christakis and Fowler 2008; Aral and Walker 2012; Wagner et al. 2012], meaning that if neighbors of a node i are active they will positively influence node i to become active as well. The magnitude of the influence, or the "speed" at which the influence is transferred from an active node to its neighbors will depend on two quantities (cf. Figure 2):

(i) *Critical Activity Threshold* a_c, which represents a soft threshold of activity that marks the point when users have an activity potential, that notably exercises influence on their peers. Note that influence is exercised at all levels of a_c. However, once a_c is reached, the influence is determined as "notable" (e.g., a level of activity that is above the average activity per user) for the corresponding peers. Hence, this critical level of activity is a system-dependent quantity. One can imagine that in a system with high user activity (e.g., a large number of changes per user) the critical activity is higher than in a system with lower levels of activity. For example, in the latter case the users will sooner notice a neighbor who became active recently. We model the *Critical Activity Threshold* as a continuous threshold. Meaning that active users will always influence their neighbors, but will exercise more influence after they have passed the critical level of activity.

(ii) *Maximum Peer Activity Flow* q represents the maximum activity flow per unit time from users to each of their neighbors. This maximum flow is reached as user activity approaches infinity. However, substantial amounts of the maximum flow are already reached whenever the user activity passes the level of the critical activity a_c.

Thus, to model peer influence, we resort to a monotonically increasing function, where more active neighbors are always more influential than less active ones. Additionally, the function $g(a_j)$ saturates for sufficiently large values of activity, inducing a natural limit on how much users can be influenced by their neighbors. We model this by setting $g(a_i, a_j) =$

$g(a_j)$ and choosing an algebraic sigmoid function with:

$$g(a_j) = \frac{qa}{\sqrt{a_c^2 + a_j^2}}, q, a_c > 0. \tag{3}$$

Peer influence can also be analyzed in terms of the growth rate of $g(a)$, in the form of the derivative dg/da of the function $g(a)$. After simplifying and rearranging, the growth rate can be calculated as:

$$\frac{dg}{da} = \frac{qa_c^2}{(a_c^2 + a^2)^{3/2}}. \tag{4}$$

In the limit of large activity a the derivative of $g(a)$ tends towards zero, thus peer influence saturates at q. On the other hand, the maximum change in influence is observed when $a = 0$—neighbors who suddenly become active will be noted most, in terms of activity, by their peers.

2.1. Dynamics Equation

With $f(a_i)$ and $g(a_j)$ defined, the activity dynamics equation becomes:

$$\frac{da_i}{dt} = -\lambda a_i + \sum_j A_{ij} \frac{qa}{\sqrt{a_c^2 + a_j^2}}. \tag{5}$$

The different parameters of the equation have dimensions. For example, a_i and a_c have activity as unit, t has seconds as unit, λ is a rate and has inverse seconds as unit, and q has activity per second as unit. Further, the equation has three free parameters, which span a huge parameter space that is difficult to explore in detail. Therefore, our first step is to simplify the equation and express it in a dimensionless form, which typically also has a smaller number of parameters as only their relative ratios, rather than their absolute values, are of importance. Another advantageous side-effect of a dimensionless formulation is that it eliminates the absolute values of the properties under investigation, in our case user activity, which can be difficult to interpret.

There are many ways to eliminate dimensions from such equations [Lin and Segel 1988]. A useful heuristic is to try to first eliminate the dimensions from the most non-linear term in the equation, which in our case is $g(a_j)$. Thus, we begin by defining a relative activity x as the ratio between the activity a and the critical activity a_c:

$$x = \frac{a}{a_c}. \tag{6}$$

The variable x is dimensionless now, and it is easy to interpret. For example, the fact that $x = 5$ means that users exercises a strong influence on their neighbors, since the level of activity is five times the critical activity a_c. In fact, the influence in this case is $g(5a_c) = (5q)/\sqrt{26} \approx 0.98q$. On the other hand if $x \ll 1$ (e.g., $x = 0.1$), this then means that the influence of users on their neighbors is much smaller as $g(0.1a_c) = (0.1q)/\sqrt{1.01} \approx 0.1q$.

By rearranging, substituting x for a and simplifying (a_c cancels in the second term) our activity dynamics equation reduces to:

$$a_c \frac{dx_i}{dt} = -\lambda a_c x_i + \sum_j A_{ij} \frac{qx_j}{\sqrt{1 + x_j^2}}. \tag{7}$$

To eliminate the dimensions from the second term we divide both sides with q:

$$\frac{a_c}{q} \frac{dx_i}{dt} = -\lambda \frac{a_c}{q} x_i + \sum_j A_{ij} \frac{x_j}{\sqrt{1 + x_j^2}}. \tag{8}$$

The term q/a_c is the growth rate of the function $g(a)$ evaluated at zero:

$$\left.\frac{dg}{da}\right|_{a=0} = \left.\frac{qa_c^2}{(a_c^2+a^2)^{3/2}}\right|_{a=0} = \frac{q}{a_c}. \tag{9}$$

This quantity gives the rate at which the influence on the peers grows if the user activity experiences a small displacement from the point of zero activity. Let us now define this quantity as *Peer Influence Growth Rate* and denote it with $\mu = q/a_c$ since this will simplify the algebra and will make the model interpretation more intuitive. Thus, the last equation can then be written as:

$$\frac{1}{\mu}\frac{dx_i}{dt} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}}. \tag{10}$$

Finally, we also want to scale time t and express the equation in terms of dimensionless time τ. This last reformulation will further simplify the equation and allows us to interpret and compare activity dynamics over time across various systems. The latter is possible due to the usage of dimensionless time τ to scale and compare the time evolution of different systems relative to each other. Let us make the following substitution:

$$\tau = \mu t. \tag{11}$$

By substituting τ for t in the term on the left hand side in Equation 10 we arrive at the dimensionless dynamics equation:

$$\frac{dx_i}{d\tau} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}}. \tag{12}$$

Now, there is only one parameter in our dynamics equation, namely the ratio λ/μ. This is a dimensionless ratio of two rates: (i) The *Activity Decay Rate* λ, which is the rate at which a user loses activity, and (ii) the *Peer Influence Growth Rate* μ, which is the rate at which a user gains activity due to the influence of a *single* neighbor.

The ratio between those two rates is the ratio of how much faster users lose activity due to the decay of intrinsic activity (or interest) than they can gain due to positive peer influence of a single neighbor. For example, a ratio of $\lambda/\mu = 100$ would mean that the users intrinsically lose activity 100 times faster than they potentially can get back from one of their neighbors. If we would set $\lambda/\mu = 1$, it would mean that users would lose activity as fast as they can regain it from one of their peers. For a short description of all parameters of the activity dynamics model see Table I.

2.2. Linear Stability Analysis

In general, Equation 12 is a coupled set of n (n being the number of nodes or users in the network) non-linear differential equations, for which, in a typical case, no closed form solution can be found. Therefore, we turn our attention to the properties of so-called fixed points. A fixed point $\boldsymbol{x^*}$ represents all the values for x_i^* for which the system does not change in time:

$$\frac{dx_i}{d\tau} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}} = 0, \forall i. \tag{13}$$

Suppose that we are able to find a fixed point $\boldsymbol{x^*}$ by solving Equation 13. One obvious fixed point in our model is $\boldsymbol{x^*} = \mathbf{0}$, meaning that x_i^* has the same value for every i: $x_i^* = x^* = 0$, representing a simple special case: a symmetric fixed point. We can easily check that $x^* = 0$ is indeed a fixed point since $f(x^*) = g(x^*) = 0$, and this also gives $f(x^*) + \sum_j A_{ij} g(x^*) = 0, \forall i$.

We are investigating this specific fixed point, as it also has a particular interpretation in our model. At this fixed point all users have zero activity, which means that they are completely inactive and the system is in an inactive or "dead" state. If the system is in such a state and no external incentives are provided, nothing will ever change and the system will remain inactive indefinitely.

Typically, we are interested in the implications on the system if we provide a small enough impulse to leave such a steady (inactive) state. In our context, the most interesting question is if the system will move from an inactive state towards a state of lively activity or if it will just revert to the inactive state. Technically, we are interested in the stability of the fixed point. In particular, we want to know if the fixed point is attracting (meaning that the system's activity in the proximity of the fixed point will be attracted to it) or repelling (meaning that the system's activity close to the fixed point will be pushed away from it).

To answer this question we linearize the functions in the proximity of a fixed point. We represent the value of x_i close to the fixed point with $x_i = x^* + \epsilon_i$, where ϵ_i is sufficiently small. To simplify the calculations, we concentrate on the case of a symmetric fixed point, such as $\boldsymbol{x^*} = \mathbf{0}$. Next, we perform a Taylor expansion about the fixed point and linearize by neglecting the terms of second and higher orders. After simplification we obtain (for details see e.g. Newman [2010]):

$$\frac{d\epsilon_i}{d\tau} = -\frac{\lambda}{\mu}\epsilon_i + \sum_j A_{ij}\epsilon_j, \tag{14}$$

where ϵ_i is the displacement of x_i from the fixed point x^*.

We can also write Equation 14 in matrix form, which gives:

$$\frac{d\boldsymbol{\epsilon}}{d\tau} = (-\frac{\lambda}{\mu}\boldsymbol{I} + \boldsymbol{A})\boldsymbol{\epsilon}, \tag{15}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{A}$ is the adjacency matrix.

Table I: **Model and model parameters.** The activity dynamics equation is in a dimensionless form and scales over relative time τ. All properties, as well as the single parameter of the model, are briefly described under *Properties* and *Parameters*.

Equation	Name
$\frac{dx_i}{d\tau} = -\frac{\lambda}{\mu}x_i + \sum_j A_{ij}\frac{x_j}{\sqrt{1+x_j^2}}$	Activity Dynamics Equation
Properties	**Name**
λ	*Activity Decay Rate*
q	*Maximum Peer Activity Flow*
a_c	*Critical Activity Threshold*
$\mu = \frac{q}{a_c}$	*Peer Influence Growth Rate*
τ	Relative Time Scale
Parameter	**Name**
$\frac{\lambda}{\mu}$	The ratio, describing how fast users intrinsically loses activity compared to how fast they get it back from (one of) their neighbors.

We can solve the last equation by writing $\boldsymbol{\epsilon}$ as a linear combination of eigenvectors $\boldsymbol{v}_r$ of the symmetric real matrix $(-(\lambda/\mu)\boldsymbol{I} + \boldsymbol{A})$:

$$\boldsymbol{\epsilon}(\tau) = \sum_r c_r(\tau)\boldsymbol{v}_r. \tag{16}$$

Equation 15 then becomes:

$$\sum_r \frac{dc_r}{d\tau}\boldsymbol{v}_r = (-\frac{\lambda}{\mu}\boldsymbol{I} + \boldsymbol{A})\sum_r c_r(\tau)\boldsymbol{v}_r = \sum_r c_r(\tau)(-\frac{\lambda}{\mu} + \kappa_r)\boldsymbol{v}_r, \tag{17}$$

where κ_r are the eigenvalues of the graph adjacency matrix $\boldsymbol{A}$. We also used the fact that the matrix $(-(\lambda/\mu)\boldsymbol{I} + \boldsymbol{A})$ has the same eigenvectors as $\boldsymbol{A}$, but with the eigenvalues $-\lambda/\mu + \kappa_r$.

The solution of the last equation for the coefficients of the linear combination is then:

$$\frac{dc_r}{d\tau} = (-\frac{\lambda}{\mu} + \kappa_r)c_r(\tau) \implies c_r(\tau) = c_r(t_0)e^{(-\frac{\lambda}{\mu}+\kappa_r)\tau}. \tag{18}$$

Now, the displacement from the fixed point will decay in time towards 0 if the exponents for the coefficients $c_r(\tau)$ are all negative. Thus, we arrive at the master stability equation for the special case of a dynamical system that we defined as:

$$-\frac{\lambda}{\mu} + \kappa_r < 0, \forall r, \tag{19}$$

Since the adjacency matrix has both positive and negative eigenvalues, a necessary stability condition is $\lambda/\mu > 0$, which is satisfied by definition. Thus, we can rearrange Equation 19 and obtain the following inequality:

$$\kappa_1 < \frac{\lambda}{\mu}. \tag{20}$$

where κ_1 is the largest positive eigenvalue of the graph adjacency matrix. Note that this inequality separates the network structure (κ_1) from the activity dynamics (λ/μ).

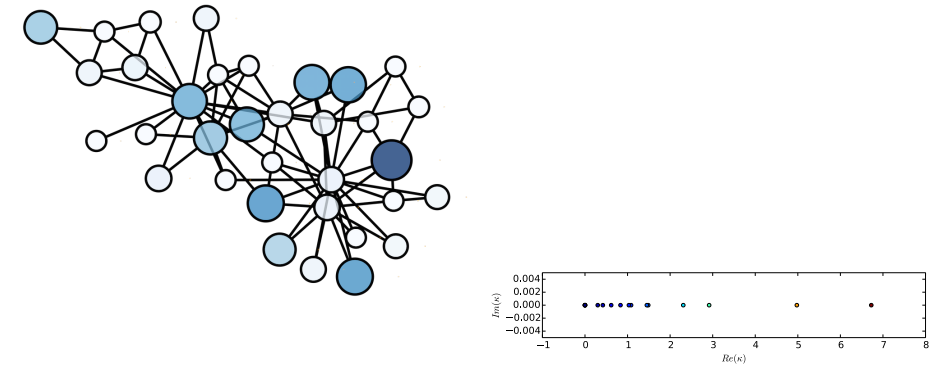
If this stability condition is satisfied, the fixed point $x^* = 0$, in which there is no activity at all ("inactive" system), represents a stable fixed point. This also means that small changes in activity only cause the system to momentarily leave the (attracting) fixed point until it becomes inactive again.

For illustration, we initialized Zachary's Karate Club Network (cf. Figures 3(a) and 3(b)) with random activities between 0 and 0.1 per node and simulate activity with our model. If the master stability equation holds (Figure 3(c)), activity converges towards zero. However, when invalidating the master stability equation (Figure 3(d)), activity converges to a new and permanently active fixed point.

In practice, additional system configurations are imaginable. Whenever the ratio is below κ_1, the system becomes unstable leaving the inactive state. However, due to the special form of the peer influence function, which saturates for large values of activity, the system will converge towards another stable state of immanent activity (i.e., ratios for periods $1-5$ of Figure 4).

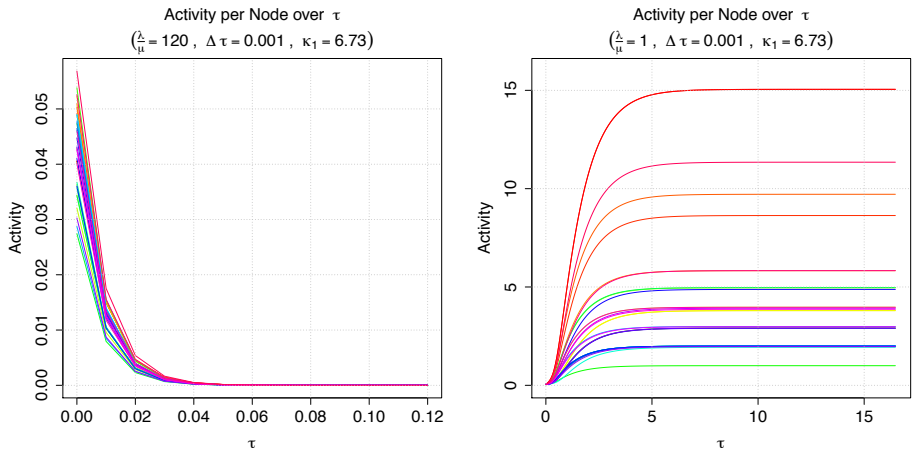
Thus, if the system is in the state where $\kappa_1 > \lambda/\mu$, we can think of **three different activity evolution scenarios**, depending on the current levels of activity present in the network:

(1) If the levels of activity are lower than the ones the network converges towards with the new ratio, we will see an increase in activity (e.g., timespans $1-2$ of *Activity Increase* in Figure 4).



(a) Zachary's Karate Club Network

(b) Adjacency Spectrum ($\kappa_1 = 6.726$)



(c) Activity Evolution $\frac{\lambda}{\mu} > \kappa_1$

(d) Activity Evolution $\frac{\lambda}{\mu} < \kappa_1$

Fig. 3: **Illustrative example. Top Left (a):** Visualization of Zachary's Karate Club. The size and color of a node represent random activity values between 0.0 and 0.1 of the corresponding nodes (bigger and darker equals higher values). **Top Right (b):** Eigenvalue spectrum of Zachary's Karate Club network. The highest eigenvalue is 6.726. **Bottom (c and d):** Evolution of activity with random initial activities (averaged over 10 runs). **Bottom Left (c):** Activity dynamics with parameters satisfying the master stability condition $\kappa_1 < \lambda/\mu$. Each line represents one node; all activities converge to the state of zero activity. **Bottom Right (d):** Invalidation of the master stability condition $\kappa_1 < \lambda/\mu$, activity converges towards a new and permanently active fixed point.

(2) If the new ratio lets the system converge towards lower levels of activity than currently present, activity will decrease, even though $\kappa_1 > \lambda/\mu$ (e.g., see timespans $2-3$ or $4-5$ of *Activity Variation* and *Activity Decrease* of Figure 4).

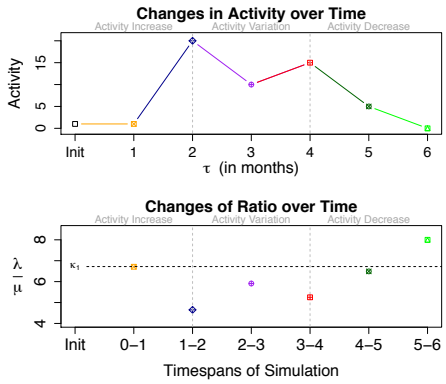


Fig. 4: **Coupled evolution of activity and λ/μ.** The **top** Figure depicts the evolution of activity (y-axis) over time (x-axis; in months) for Zachary's Karate Club network with synthetically created (random) activities. The ratios, which correspond to the activity evolutions over time in the top Figure, are depicted in the **bottom** Figure (same symbol and color), with the y-axis representing the value of the ratio, while the different timespans are depicted on the x-axis. As long as $\lambda/\mu < \kappa_1$ the network converges towards a state of immanent activity, yet decreases in activity are possible (see timespans $2-4$ of*Activity Variation* sections in **top** and **bottom**). If $\lambda/\mu > \kappa_1$ the network converges towards an inactive state.

(3) Lastly, the levels of activity have already converged towards their fixed point and λ/μ is left unchanged, retaining the levels of activity from the past (e.g., see timespans $0-1$ of *Activity Increase* in Figure 4).

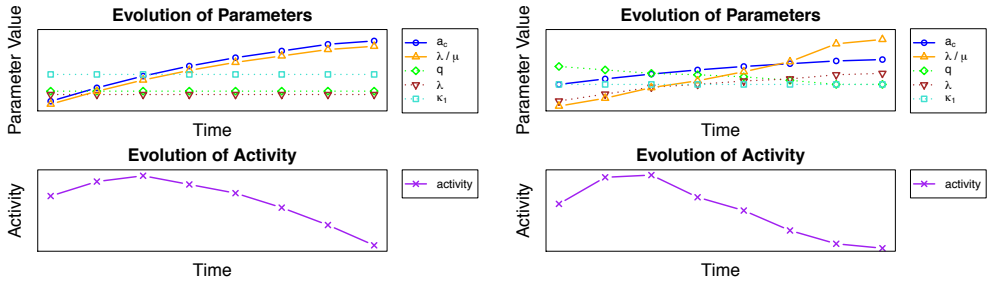
If $\kappa_1 < \lambda/\mu$ holds, the system is stable and activity converges towards the attracting fixed point at zero activity (see timespans $5-6$ of *Activity Decrease* in Figure 4).

Summary of system stability analysis. In order to permanently leave the stable state of complete inactivity we are interested in making the system unstable. To be able to leave the attracting force of the fixed point at zero activity we have the following two options:

(i) We **provide (continuous) external impulses** to the system, for example, in the form of incentives for users to increase their activity, pushing the system far away from the fixed point of no activity (and hope that it will be attracted by another fixed point where activity is not zero).
(ii) We **compromise the stability condition** by either manipulating:
  (a) the network structure (i.e., *making κ_1 larger*) or
  (b) the activity dynamics (i.e., *making λ/μ smaller*).
  Structurally, we can manipulate the size of κ_1 by creating or removing links (and nodes) in our network (for more information on how to manipulate κ_1 see [Newman 2010]). Dynamically, λ/μ becomes smaller if either λ becomes smaller, meaning that the intrinsic user activity decays at a slower pace or μ becomes larger, meaning that people copy their friends more and faster, or both.

2.3. Discussion on Parameter Evolution

At this time, we leave the investigation of the manipulation of the activity dynamics ratio λ/μ as well as the manipulation of the network structure to invalidate the master stability equation open for future work. Nevertheless, before illustrating how our proposed activity



(a) Evolution of *Critical Activity Threshold* (b) Coupled Evolution of Parameters

Fig. 5: **Parameter Evolution Scenarios.** In a system with (at first) increasing overall levels of activity and fixed values for q and λ for all users, we expect a_c to slowly increase (see **(a)**), as individual contributions are indistinguishable due to a flood of newly added content (activity). As a consequence, more posts and replies are required from all users to exercise the same amount of peer influence—represented by increasing values for a_c over time. After a certain point in time, a_c will reach a threshold and activity will start to decrease, if not intervened by administrators. In a more realistic scenario (see **(b)**), again with increasing levels of overall activity, users will—in addition to increasing values of a_c—start to lose interest in contributing to the system, represented by increasing values for λ. As a consequence, activity will decrease at a faster pace.

dynamics model can be applied to empirical datasets, we discuss potential system evolution scenarios and their implications for activity.

Activity Decay Rate. Technically, if λ increases, the ratio λ/μ increases as well, resulting in higher (faster) losses of activity per timespan. Once the system satisfies the master stability equation ($\kappa_1 < \lambda/\mu$) it will inevitably become inactive. To be precise, the larger λ for a stable system, the faster activity will converge towards zero. Essentially, an increase in λ represents an increased intrinsic loss of activity for all users (e.g., due to a lack of interest to contribute) while a decrease of λ can be interpreted as an increase of interest (more precisely, slower loss of interest) and thus higher levels of activity.

Evolution scenarios of Activity Decay Rate. We would expect to see an increase in λ on websites with low levels of user interaction and activity (i.e., meaning that individual contributions are not valued, as no feedback is provided). On the other hand, websites that engage with their users and provide steady updates (e.g., new content or functionality) will likely see a consistent or even decreasing λ. In general, practitioners can influence λ by, for example, providing incentives for users to contribute, such as badges, barn stars, likes, reputation systems, or monetary incentives.

Peer Influence Growth Rate. With increasing values for μ the ratio λ/μ decreases, resulting either (i) in an overall increase in activity if the system is unstable ($\kappa_1 > \lambda/\mu$), (ii) in prolonged timespans of activity before converging towards inactivity if the system is stable ($\kappa_1 < \lambda/\mu$), (iii) or in an invalidation of the master stability equation if λ/μ reaches a tipping point where $\kappa_1 > \lambda/\mu$.

The evolution of μ directly corresponds to the evolution of the *Maximum Peer Activity Flow* and *Critical Activity Threshold*.

Maximum Peer Activity Flow. The parameter q defines the maximum amount of activity (peer influence) that can traverse along the edges of the collaboration network per unit time. If this parameter increases, $\mu = q/a_c$ will increase as well; resulting in an overall increase in activity. In contrast, reducing the value of q results in overall decreasing levels of activity.

Evolution scenarios of Maximum Peer Activity Flow. In real-world systems, q is best interpreted as a proxy for the efficiency of the user interface, describing how well information (or influence) is transported (e.g., highlighted or visualized) across users. For example, practitioners can influence the *Maximum Peer Activity Flow* by adding recommendations for users to collaborate with or by optimizing the presentation of newly added/edited content. Note that with increasing numbers of users and levels of activity it becomes increasingly difficult for practitioners to keep q at its current level, let alone positively influence the parameter due to the vast amount of content and/or activity present in the system.

Critical Activity Threshold. The parameter a_c represents a soft threshold, which defines when users start to "effectively notice" the actions of their peers and are, as a consequence, "notably" influenced (see Figure 2(b)) by them. The larger a_c, the more actions (i.e., posts or replies) are required by users to positively influence their peers to copy their actions and increase their activity levels (see Figure 5).

Evolution scenarios of Critical Activity Threshold. In practice, we would expect to see an increasing a_c with an increasing number of active users and levels of activity. For example, in a system with low activity and a small number of users, each action by a particular user will be noticed immediately by all others—meaning that the level of a_c is low. However, with increasing numbers of users and an increase in activity, users have to increase their number of posts and replies to be noticed by their peers. Hence, the more active users are present in a system, the harder it becomes for users to specifically notice each contribution of their peers individually. In a worst case, users are confronted with an activity overload that might even result in decreasing levels of (positive) peer influence. In particular, an initial increase in activity likely leads to an increase in a_c, which in turn decreases activity in the system. Thus, evolution of a_c represents a *negative feedback loop* in the system. In contrast to q, which serves as a proxy for the user-interface, a_c represents an intrinsic parameter of the users of a system. Administrators of such networks and websites can influence a_c by either influencing q (e.g., by adjusting the user interface to better promote each individual action taken by the peers of a user) or by actively avoiding and counteracting the activity overflow by filtering and reducing the amount of new content that is displayed at once.

For example, the mechanisms of how Facebook displays posts in its "News Feed" can be seen as a measure to filter and limit newly added content; actively avoiding information or activity overloads while maximizing the (peer) influence of each individual contribution.

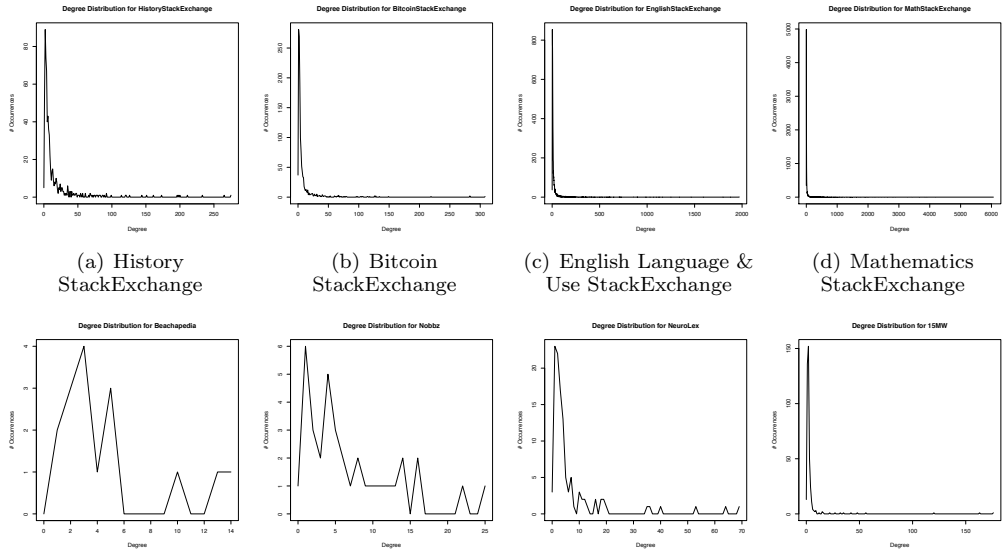
Summary of evolution scenarios. If activity increases over time and no adaptations to the system are implemented, activity will inevitably decrease, due to a larger *Critical Activity Threshold* (see Figure 5). To counteract this development, website administrator could either try to manipulate *Activity Decay Rate*—an intrinsic property that varies per user—or optimize the user interface, and thus manipulate *Maximum Peer Activity Flow*.

3. EMPIRICAL ILLUSTRATION

We are now interested in modeling and simulating activity dynamics for empirical datasets. In particular, we investigate activity dynamics for an array of different websites, consisting of instances of the StackExchange[1] network as well as multiple Semantic MediaWikis[2].

First, we characterize the investigated datasets and outline our methods for the empirical estimation of the required parameters (see Table I). We then fit our model to the collaboration networks and present the results of the activity dynamics simulation.

[1] http://www.stackexchange.org/sites
[2] http://www.semantic-mediawiki.org



(a) History StackExchange (b) Bitcoin StackExchange (c) English Language & Use StackExchange (d) Mathematics StackExchange

(e) Beachapedia Wiki (f) Nobbz Wiki (g) NeuroLex Wiki (h) 15Mpedia Wiki

Fig. 6: **Degree Distribution of Empirical Collaboration Networks.** Visualization of the degree distribution of all investigated collaboration networks. The **top row (a to d)** depicts the different StackExchange collaboration networks, while the **bottom row (e to h)** shows the collaboration network visualizations for the different Semantic MediaWiki instances. The majority of users, across all collaboration networks, exhibits between 0 and 10 collaboration edges.

3.1. Datasets

We selected a total of four differently sized instances from the StackExchange network as well as four different Semantic MediaWiki instances to model activity dynamics. In particular, we concentrate our efforts on the History StackExchange[3] (HSE), which is the smallest of the StackExchange datasets and allows users to discuss topics and questions related to history and historical events. The Bitcoin StackExchange[4] (BSE) as well as the The English Language & Usage StackExchange[5] (ESE) represent two medium-sized websites and are platforms for asking and discussing questions related to everything related to mining, buying and selling of bitcoins and the English language respectively. On the Mathematics StackExchange[6] (MATHSE) website, which also represents our largest dataset, users can ask and discuss mathematics related questions and topics.

We further investigate activity dynamics for the Beachapedia Wiki[7] (BP), representing the smallest dataset in our activity dynamics analysis, striving to create a structured knowledge base for a variety of topics on beaches in the United States. The medium-sized german Nobbz Wiki[8] (NZ) provides a structured knowledge base and discussion platform

[3]http://history.stackexchange.com
[4]http://bitcoin.stackexchange.com
[5]http://english.stackexchange.com
[6]http://mathematics.stackexchange.com
[7]http://www.beachapedia.org
[8]http://nobbz.de/wiki

for the online game "Die Verdammten"[9]. The second largest dataset, the NeuroLex Wiki[10] (NLX), represents a large and semantically enriched lexicon on terms and topics related to neuroscience. Our largest dataset is the 15Mpedia Wiki[11] (15MW)—a Spanish Semantic MediaWiki instance that discusses a wide variety of topics related to Spain and its different areas and regions.

In general, the investigated datasets are very diverse in their characteristics, for example, the number of active users ranges from $35,476$ in MATHSE to a total of 16 in BP. For the analyses conducted in this paper we focus on the last 52 weeks of each dataset. For more detailed information see Table II. The different degree distributions for all collaboration networks are highly heterogeneous (cf. Figure 6). For all investigated datasets, the majority of users exhibit between 0 and 10 collaboration edges. However, in all datasets there are a few users with a large number of collaboration edges.

From each of these datasets we extracted a collaboration network for the tasks of fitting the model and simulating activity dynamics. Hence, we first parsed the change-logs of all datasets. Each user, who has contributed at least one question, answer or comment for the StackExchange datasets, or created or edited an article for the Semantic MediaWikis is represented as a node in the corresponding collaboration network. Edges between users represent collaboration and are undirected. For the StackExchange datasets, we defined collaboration as either posting an answer to a question or posting a comment on the initial question or an answer. For the Semantic MediaWiki instances, we have created an edge between users who (chronologically and) successively changed the same article (cf. Figure 7). Edges with the same source and target user have been removed in all datasets.

Further, users with zero collaboration edges are initialized analogously to all other users and are not specifically filtered from our datasets. However, due to missing positive peer influence, activity will inevitably—as long as $\lambda/\mu > 0$—converge towards zero for these users.

Note that the presented approach for creating collaboration networks represents just one of many different possibilities to create such networks and is analogous to (undirected) co-authorship networks as presented in Newman [2001]; Barabâsi et al. [2002]. Given that the created collaboration networks are based on interactions between users, we assume similar characteristics to social networks, particularly with regards to potential peer influence [Aral et al. 2009].

[9]http://www.dieverdammten.de/
[10]http://neurolex.org/
[11]http://wiki.15m.cc/wiki/Portada

Table II: **Dataset statistics.** Note that all datasets differ in the number of users, collaboration edges and activity. Users refers to the number of unique users that have contributed more than one post or reply to the corresponding datasets within our observation periods. Posts represent newly created questions in the case of the StackExchange network and newly created articles in the case of the Semantic MediaWiki datasets. Replies are either comments or answers for all StackExchange datasets and edits of existing articles for Semantic MediaWikis. κ_1 denotes the largest eigenvalue of the corresponding collaboration network. For our experiments we limited our observation periods to the last $52 + 3$ weeks of each dataset.

Dataset	HSE	BSE	ESE	MATHSE	BP	NZ	NLX	15MW
Users	682	1,299	7,893	35,476	16	36	112	394
Edges	5,179	5,528	83,457	477,133	38	125	383	772
κ_1	54.33	43.88	162.04	303.58	6.71	11.46	18.4	19.97
Posts & Replies	12,496	12,295	151,028	986,996	2,718	603	33,792	102,521
Weeks	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$	$52+3$

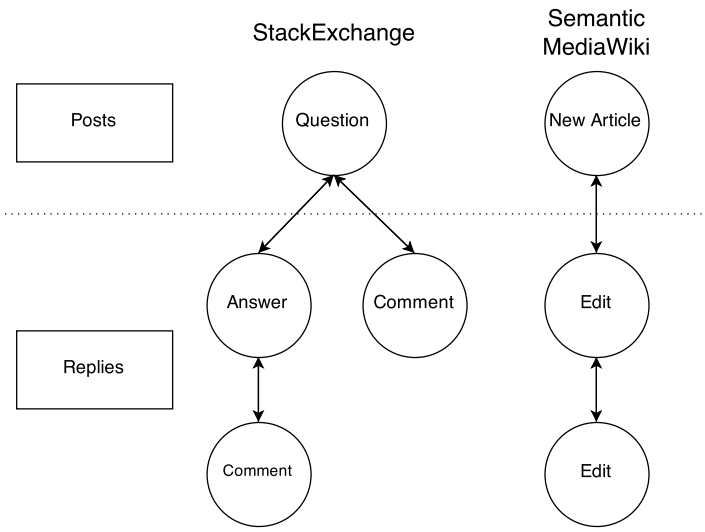


Fig. 7: **Collaboration Network Construction.** This plot depicts the different elements of the StackExchange and Semantic MediaWiki datasets that have been classified as posts and replies (cf. Table II) as well as the edges that have been drawn between certain entities and change-actions and represent collaboration in our collaboration networks.

3.2. Parameter Estimation with Least-Squares

To estimate λ/μ for (preprocessed) empirical datasets we resort to an output-error estimation method. First, we formulate the estimation of the model parameter as an optimization problem. As objective function we use a well-known least-squares cost function. Second, we solve the optimization problem numerically, using the method of gradient descent in combination with Newton's method to speed up the calculations. Finally (as a proof of concept), we evaluate the accuracy of the ratio estimate by calculating prediction errors on unseen data. Next, we describe these estimation steps in more details.

Preprocessing. First, we aggregate all activities per user per day and apply a rolling mean of 7 days to smoothen and reduce strong fluctuations in activity, which are likely caused by external influences. Second, we further aggregate the smoothed activities per user and per (calendar) week. For an additional noise reduction in our datasets we remove all users that have contributed less than one post or reply in the smoothened dataset during our observation period, as well as the first and last week of our datasets, if they contain less than 7 days of activity data. Finally, since we only want to illustrate the practical application of our model on the empirical data we extract the last $52+3$ to weeks from all our datasets. Note that the 3 additional weeks are required to calculate a ratio for the simulation of activity for the first week.

Formulating estimation as an optimization problem. Depending on a particular application of the model we may need to introduce a suitable objective function. For example, we may be interested in applying our model to analyze and simulate the aggregated levels of activity in a system. In other words, we are interested in the overall activity level in a system, rather than in the particular activity distribution over the users (see below for another example involving user activity levels). Hence, we formulate the objective function (see Equation 21) as a least squares cost function, which calculates the error of the sum of activity over multiple data points over a certain period of time T:

$$J(\frac{\lambda}{\mu}) = \frac{1}{T}\sum_{k=0}^{T-1}\left[\sum_{i}^{n} x_i(k+1) - \sum_{i}^{n}\hat{x}_i(k+1)\right]^2, \tag{21}$$

where $x_i(k)$ is the empirically observed activity of user i at time k, $\hat{x}_i(k)$ is the estimated activity for user i at time k, and n is the total number of users as before.

To calculate the estimates $\hat{x}_i(k)$ we numerically integrate the differential equations from our model by applying Euler's method for solving differential equations computationally. Thus, we approximate the time evolution of $\hat{x}_i$ between all time steps k and $k+1$ (for each of these steps we set the total time to τ) by iterating:

$$\hat{x}_{i,t+1}(k) = \hat{x}_{i,t}(k) + \Delta\tau\left[-\frac{\hat{\lambda}}{\mu}\hat{x}_{i,t}(k) + \sum_j A_{ij}\frac{\hat{x}_{j,t}(k)}{\sqrt{1+\hat{x}_{j,t}(k)^2}}\right], \tag{22}$$

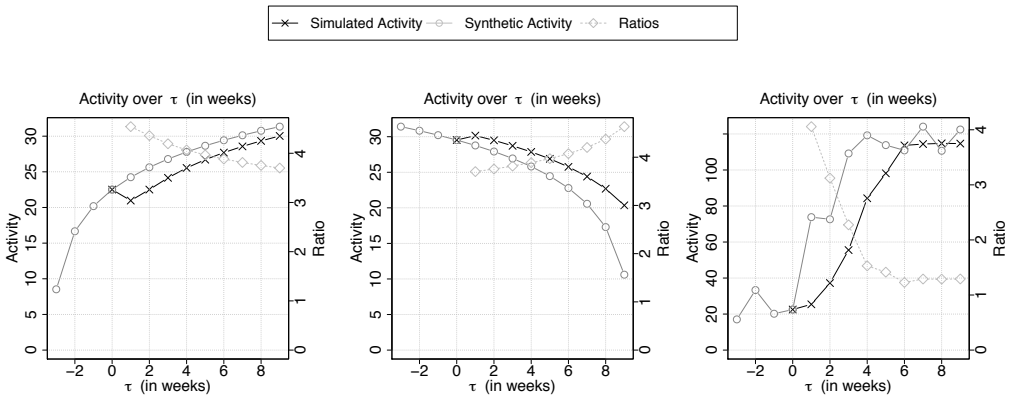
where we set $\hat{x}_{i,t=0}(k) = \hat{x}_i(k)$, $\forall i,k$ and use the current estimate for λ/μ to perform calculations. The final equation for $\hat{x}_i(k+1)$ becomes:

$$\hat{x}_i(k+1) = \hat{x}_i(k) + \Delta\tau\sum_{t=0}^{t=\tau}\left[-\frac{\lambda}{\mu}\hat{x}_{i,t}(k) + \sum_j A_{ij}\frac{\hat{x}_{j,t}(k)}{\sqrt{1+\hat{x}_{j,t}(k)^2}}\right]. \tag{23}$$

The local approximation error for the Euler's method is of the order $O(\Delta\tau^2)$ and the global of the order $O(\Delta\tau)$. To perform integration between steps k and $k+1$ we need to iterate for $\tau/\Delta\tau$ steps, where $\Delta\tau$ needs to be chosen with care. In general, if we set $\Delta\tau$ too high—meaning that the calculations are less computationally intensive, as we have to run a smaller number of iterations—the accuracy of our simulation (including the estimation of the ratio) will decline, as the potential error per iteration due to our approximations becomes higher. This error can become so large that it could potentially lead to numerical instability, meaning that the overall activity in a system can become negative, which might result in activity to diverge towards $\pm\infty$. With certain combinations of the network structure, $\Delta\tau$ and the calculated ratios, activity can become negative without diverging, oscillating around the fixed point of zero activity until convergence. In contrast, if we set $\Delta\tau$ too low we end up with a very precise simulation, although the time necessary to compute the simulation will be much higher, as a much larger number of iterations will have to be executed.

Numerical solution of the optimization problem. We solve the optimization problem numerically using the method of gradient descent. The first derivative of the objective function (Equation 21) defines the update rule or gradient, which directs if and to what extent we have to increase or decrease λ/μ to minimize the error of the sum of activities over all data points during T.

Once we calculate the first derivative with the current values of estimated activities we update the ratio by multiplying the derivative with the *learning rate* η. Thus, the complete procedure is as follows. First, we initialize our estimation by using κ_1 for the first simulation. Second, we estimate the activities and calculate the gradient with these estimates. Third, we calculate the error between our simulated and empirical values, and adapt the ratio according to the corresponding update function and step size η. Fourth, we repeat this process until the calculated update for the ratio is smaller than a given convergence criterion (e.g., 10^{-12}) or if we reach a total of $20,000$ iterations without reaching convergence. Additionally, we have also implemented Newton's method, which in our cases substantially reduces the computation time. In all our experiments we set T to four weeks, meaning that we optimize the objective function by calculating the optimal ratio over a span of four data points (weeks).



(a) Increasing Activity (b) Decreasing Activity (c) Variable Activity

Fig. 8: **Illustrations with Synthetic Data.** The plots depict the results of the activity dynamics simulations for Zachary's Karate Club network with synthetic activity values (left y-axes) and the corresponding ratios (right y-axes). The black solid lines with x markers represent the simulated activity over t (in weeks; x-axes). The solid gray lines with circles represent synthetic activities; the gray dotted lines with diamonds represent the ratios corresponding to the simulated activities. With increasing and decreasing activities, the ratios become smaller (see **(a)**) and larger (see **(b)**). When setting activity randomly (see **(c)**) the ratio adjusts analogously.

Evaluation of the parameter estimates. We evaluate the accuracy of the estimated parameters by cross-validation (leave-one-out method). In particular, we use the estimated ratios over 4 weeks to simulate activity for the succeeding week. For example, we calculate the optimal λ/μ (according to our objective function) for weeks 1 – 4 and predict activity for week 5. Next, we use the empirical data of weeks 2 – 5 to calculate the ratio to predict activity for week 6. Hence, we calculate a total of 52 ratios to simulate activity for a total of 52 weeks.

As depicted in Figure 8, we have created three synthetic scenarios to test and illustrate the mechanisms of the *Activity Dynamics Model*. First, we estimate λ/μ (right y-axes; gray dotted lines with diamonds) for the three scenarios with synthetically created increasing, decreasing and variable or random activities (left y-axes; gray solid lines with circles) over $10+3$ weeks (x-axes). In all three scenarios we use Zachary's Karate Club as the underlying collaboration network. Due to our parameter estimation process the simulated levels of activity (left y-axes; black solid lines with x markers) exhibit a small lag when activity steadily moves into one direction (i.e., increases or decreases). On the other hand, small fluctuations (see weeks 6 – 9 in Figure 8(c)) are mitigated. The ratios (right y-axes), which correspond to the simulated levels of activity in the same week, are depicted as well.

Discussion on parameter estimation method. To validate the correctness of our implementation of the method of least squares, we have simulated activity for datasets with a preset ratio (and random weights for initialization) for 3 weeks. We then used the random activity initialization values, as well as the activity values for each of the 3 weeks as input for the calculation of the ratio with the method of least squares. Using this approach, we were able to estimate previously set ratios with negligibly small errors. When adding noise to the simulated activity values, the obtained ratios were less accurate accordingly.

Note that the estimation and validation method that we apply is only one of many possible methods. In this paper, we want to illustrate the general applicability of our method as well as its potential to gather new insights into the intricate dynamics of activity in online collaboration networks. We measure the accuracy of the prediction only as a general proof of concept of our model and leave further investigations of the predictive power of our method open for the future work. Following up on this notion, we now shortly discuss some alternative approaches for formulating the objective function and their implications.

Alternative objective functions. To demonstrate the versatility of our model—if we are interested in answering questions about the distribution of the activities over users—we may change the formulation of the objective function to calculate ratios that minimize the error of activity per user and per data point (see Equation 24). Note that when optimizing towards aggregated levels of activity, we obtain ratios that characterize the systems. In contrast, with the adapted objective function, we are interested in learning more about the users of such systems. The alternative objective function may be defined as follows:

$$J(\frac{\lambda}{\mu}) = \frac{1}{T}\sum_{k=0}^{T-1}[\boldsymbol{x}(k+1) - \hat{\boldsymbol{x}}(k+1)]^2, \tag{24}$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are now n-dimensional vectors storing the activities of all n users. Thus, this objective function represents the sum of squared errors calculated for each of the n users of the corresponding systems over a total of T data points.

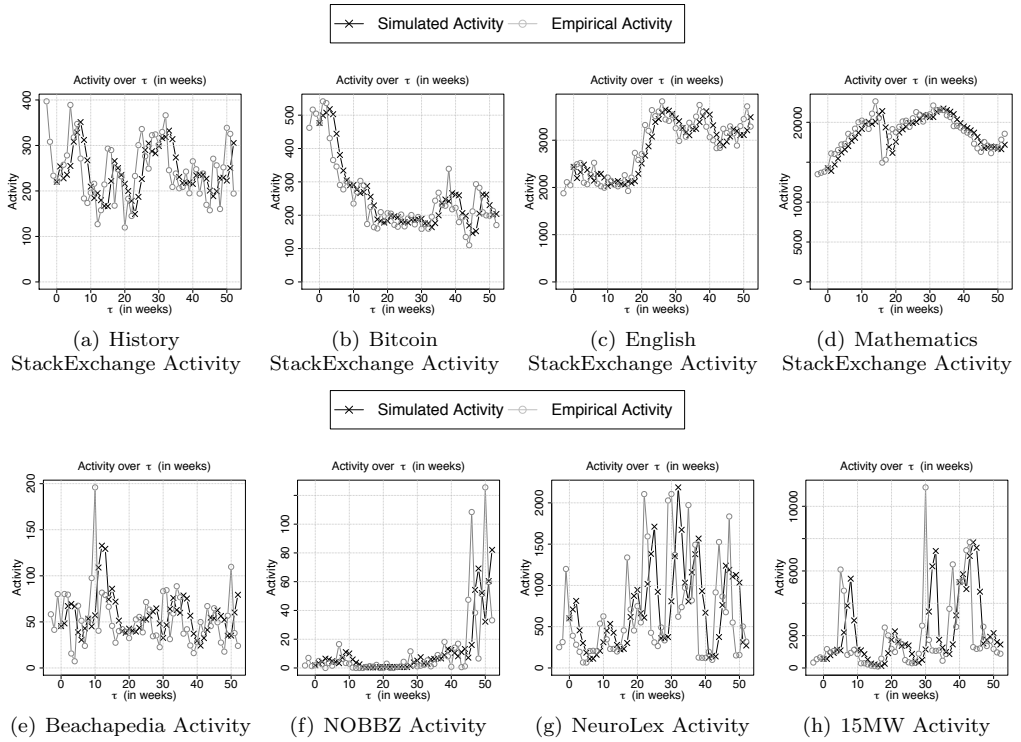
We have estimated λ/μ and simulated activity for HSE using this objective function. In contrast to the aggregated levels of activity, we obtain a more accurate distribution of activities across all users, as was intended. However, each of the 4 data points in T now corresponds to a vector of n users, as opposed to a single value (the aggregated activities), resulting in either much higher computation times, a larger error for the prediction tasks or both.

Additionally, to tackle the prediction problem and to avoid overfitting we may introduce a *regularization term* to the objective function. For example, we might be interested in keeping the ratio or the difference between the ratio and κ_1 small. In the latter case we would add a term such as $\gamma(\kappa_1 - \lambda/\mu)^2$ to our objective function, where γ represents the strength of regularization.

We leave a detailed analysis and comparison of different objective functions open for future work. The ratios calculated to minimize the error for aggregated activity levels exhibit higher accuracy in our simulations (in terms of overall activity per month). The trade-off for a more accurate distribution of activities over users with the changed objective function are worse results for the simulation of activity, as not only the aggregated activity levels are considered, but the vector of activities of all user in our datasets over multiple points in time. However, these ratios provide a better overall correlation between simulated and empirical activities per contributor of our system.

3.3. Illustration on Empirical Datasets

After calculating λ/μ and setting $\Delta\tau$ we simulate activity in our collaboration networks. Due to our chosen approximations, the main goal of the presented illustration is not to predict activity in collaboration networks. Rather, we are interested in demonstrating that our assumptions regarding the *Activity Decay Rate* and the *Peer Influence Growth Rate* hold and allow us to simulate trends in activity dynamics for given and real values. Further, by modeling and simulating activity dynamics for empirical datasets we not only deepen our understanding of the model but we also—depending on the values of the parameters—potentially obtain new insights into the systems under investigation.



(a) History StackExchange Activity (b) Bitcoin StackExchange Activity (c) English StackExchange Activity (d) Mathematics StackExchange Activity

(e) Beachapedia Activity (f) NOBBZ Activity (g) NeuroLex Activity (h) 15MW Activity

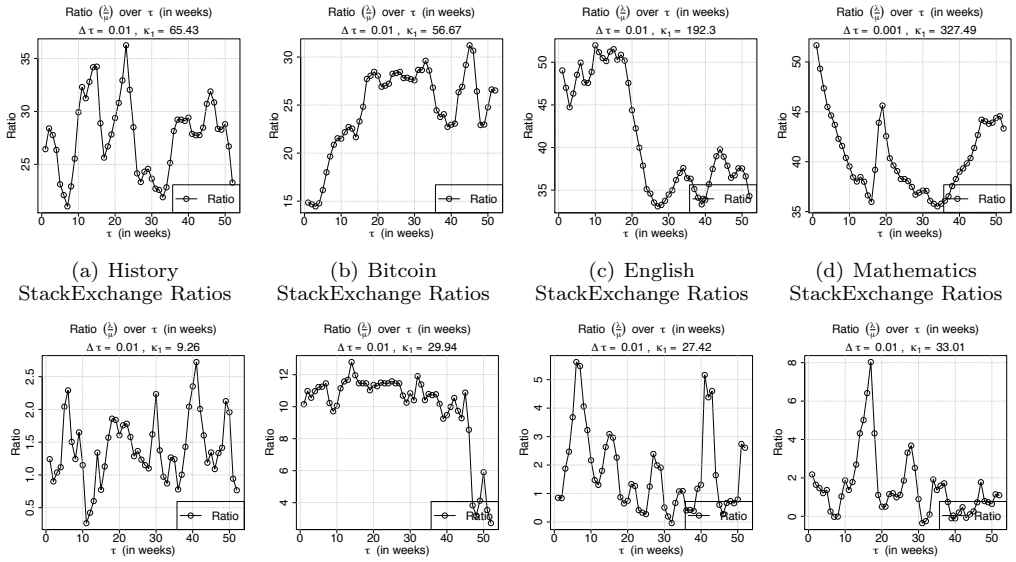
Fig. 9: **Results for the activity dynamics simulation.** The plot depicts the results of our activity dynamics simulation for the StackExchange datasets (**top row**) and Semantic MediaWiki instances (**bottom row**). The solid gray lines with circles represent the empirical (observed) activity over t (in weeks; x-axes), while the solid black lines represent the simulated activity dynamics (y-axes). In all of our analyzed datasets, the simulated activity dynamics exhibit a notable resemblance to the empirical activity.

Figure 9 depicts the results of the activity dynamics simulation. The root mean-squared errors (RMSEs) of the simulations are listed in Table III.

Overall, the results gathered from the activity dynamics simulation exhibit a notable resemblance to the real activities of the corresponding datasets. Due to the chosen approximations and simplifications when estimating λ/μ for our model (i.e., static network structure and average model parameters over weeks and users), the simulated activity is naturally limited in its accuracy. These limitations are particularly visible whenever there are large and sudden increases of activity in the collaboration networks. Note that λ/μ will

Table III: **RMSE.** The table depicts root mean-squared errors (RMSE) of our activity dynamics simulation per user and week for all datasets. Our simulation yields a small RMSE for all StackExchange datasets. RMSE for the Semantic MediaWiki datasets is slightly higher, which is likely due to the lower number of active users (listed in the Users column).

Dataset	HSE	BSE	ESE	MATHSE	BP	NZ	NLX	15MW
Activity	12,496	12,295	151,028	986,996	2,718	603	33,792	102,521
Users	682	1,299	7,893	35,476	16	36	112	394
RMSE	0.076	0.031	0.029	0.030	1.755	0.274	4.397	4.043



(a) History StackExchange Ratios (b) Bitcoin StackExchange Ratios (c) English StackExchange Ratios (d) Mathematics StackExchange Ratios

(e) Beachapedia Ratios (f) NOBBZ Ratios (g) NeuroLex Ratios (h) 15MW Ratios

Fig. 10: **Evolution of ratios λ/μ.** The evolution of the ratios λ/μ (y-axes) over τ (in weeks; x-axes) for the StackExchange datasets (**top row**) and for the Semantic MediaWiki instances (**bottom row**). The smaller the ratio, the higher the levels of activity in Figure 9. Small variances in λ/μ over time indicate that activities of the systems are less influenced by the activity of single individuals than they are by peer influence.

only be higher than κ_1 if activity in our datasets is either zero or the relative difference in activity between two months is extremely high, which is never the case for our smoothed empirical datasets.

Further, the assumption of a fixed network structure of our investigated collaboration networks also (negatively) influences the obtained results of our simulation. For example, it is possible for our simulation to yield higher increases in activity (e.g., Figure 9(b)), as users might be influenced by peers, who would join the collaboration network only at a later point in time.

4. SYSTEM MASS AND ACTIVITY MOMENTUM

We can further analyze the obtained ratios and parameters of our activity dynamics simulation to broaden our understanding of the collaboration networks under investigation. Figure 10 depicts the value of the calculated ratios λ/μ (y-axis) for each week (x-axis). If the ratio is higher than κ_1 (denoted in the title of each Figure), our master stability equation holds and the system converges towards zero activity (over time). The amount of activity that is lost per iteration—and hence the speed of activity loss—is proportional to the value of the ratio and the activity already present in the network. In general, a higher ratio results in a higher and faster loss of activity.

If the ratio is smaller than κ_1, the master stability equation has been invalidated and the system will converge towards a new fixed point of immanent activity (cf. Section 2.2). If this is the case, we can observe one of three potential behaviors, which are triggered depending on the amount of activity already present in the network and the current ratio:

(i) **An increase in activity** if the new fixed point, corresponding to the new ratio, is of higher overall activity than the activity already present in the collaboration network (see $\tau = 20 - 30$ in Figures 9(d) and 10(d)). This situation emerges whenever we invalidate the master stability equation from a previously stable fixed point or if the system is already stable in a situation when the new ratio is smaller than the last estimated ratio.
(ii) **A decrease in activity** if the new fixed point is of lower overall activity than the activity already present in the collaboration network (see τ $3 - 7$ in Figures 9(b) and 10(b)). Again this may occur in two specific situations. First, if the ratio increases, so that the master stability equation is now satisfied and the system has been previously in an unstable state. Second, if the system is in an unstable state but the ratio increases slightly without satisfying the stability equation.
(iii) **No change in activity** if the new fixed point corresponding to the new ratio is of the same overall activity than the activity already present in the collaboration network (see τ $20 - 30$ in Figures 9(b) and 10(b)).

System Mass. We can now use the obtained ratios to characterize the collaboration networks and quantify their robustness in terms of their activity dynamics. Robust systems are systems with lively and high levels of activity, which are able to keep that activity even in the cases of small unfavorable changes in the dynamical parameters. Less robust systems are systems that lose their activity very quickly as a consequence of even small changes in the ratio. Thus, we calculate the standard deviation over all ratios $\sigma_{\lambda/\mu}$ over time and normalize it over κ_1—to account for the size of the collaboration networks—and refer to it as ρ—the normalized standard deviation of the ratio λ/μ (see Equation 25).

$$\rho = \frac{\sigma_{\lambda/\mu}}{\kappa_1} \tag{25}$$

The normalized standard deviation is a measure of system sensitivity and its inverse $(1/\rho)$ represents a measure of system stability or *inertia* to changes in activity. Analogously to mass in classical mechanics—which defines the inertia or resistance of being accelerated or decelerated for an object by a given force—we call the quantity $1/\rho$ the *System Mass*. We denote this quantity with m_s with the subscript s to distinguish it from the number of links m in a collaboration network (see Table IV). In systems with a large *System Mass* it is more difficult to induce changes in activity. In particular, this means that it is more difficult to reduce activity in a consistently active system (due to the small standard deviations of λ/μ), as well as it is difficult to jump-start the same system if activity levels were consistently low in the past (again, due to small standard deviations of λ/μ).

Activity Momentum. After calculating the *System Mass* m_s, we are now interested (again analogously to classical mechanics) in calculating the *Activity Momentum* p for our collaboration networks (see Equation 26).

$$p = m_s a \tag{26}$$

For activity we take (i) the average activity (posts and replies) per week and (ii) the activity in the last month of our observation periods (cf. Table IV) and calculate (i) the average and (ii) the current momentum.

The higher the *Activity Momentum* of a collaboration network, the more force is needed to "stop" (make it inactive) the system. Hence, the higher the momentum, the more robust a given network. In particular, if a (sufficiently) small number of users would suddenly stop contributing to a collaboration network that exhibits a very large *Activity Momentum* p, activity in the overall network would be minimally influenced. On the other hand, if the same number of users would stop contributing to a collaboration network with a (significantly)

smaller *Activity Momentum* p, chances are that their actions (or lack thereof) will have a notable influence on the overall trends in activity dynamics of the system. In particular, there are three factors that influence the *Activity Momentum* of collaboration networks:

(i) *The standard deviation of* λ/μ. If the ratio is very stable and does not frequently oscillate, the standard deviation and hence the normalized standard deviation will be very small. This also means that activity, as well as increases and decreases thereof, is equally distributed across τ and is not (frequently) exercised in bursts.
(ii) *The largest eigenvalue* κ_1. Larger and denser collaboration networks exhibit a larger highest eigenvalue κ_1. As ρ is the normalized variance of the ratios over κ_1, the largest eigenvalue will directly influence ρ. The notion of normalizing ρ over κ_1 follows the intuition that that large collaboration networks are less likely to exhibit sudden changes in activity than smaller ones.
(iii) *The activity.* The larger the average activity (posts and replies) per month, the higher the *Activity Momentum* of a collaboration network, and hence the higher the force that is needed to render the collaboration network inactive. Analogously, networks with a small *Activity Momentum* require less force to be influenced (i.e., to either speed up/increase or slow down/decrease activity).

Hence, we can use the calculated *Activity Momentum* p as an indicator of the activity level as well as the tendency of a system to stay at that activity level in the future. For example, MATHSE exhibits the most robust collaboration network of our datasets regarding changes in activity, with an *Activity Momentum* of order 10^6 (average per week and last month). ESE and 15MW both exhibit similar average *Activity Momenti* of orders 10^4. However, when looking at the *Activity Momenti* of the last months, ESE is roughly four times as hard to stop as 15MW.

In contrast, HSE and BSE exhibits very similar activity levels for last month, however the corresponding *Activity Momentum* of HSE is twice the one of BSE, indicating that half the force is needed to render BSE inactive than it would be needed to render HSE inactive. The other datasets follow analogously.

On the other hand, BP exhibits a high value for *System Mass* and a very low corresponding *Activity Momentum*, indicating that it will be very difficult to to accelerate or jump-start the system with regards to activity.

Table IV: **System Mass and Activity Momentum.** The table depicts the results for the activity momentum analysis. ρ is the standard deviation of the calculated ratios normalized over κ_1. System Mass is represented by $1/\rho$ and Activity Momentum represents System Mass multiplied with Activity. Activity depicts the average activity per week as well as the value for the last observed months in brackets. Activity Momentum follows analogously. MATHSE and ESE exhibit the largest average and current Activity Momenti, followed by 15MW and NLX. Even though 15MW exhibits a System Mass similar to HSE and NZ, its Activity Momentum is much larger.

Dataset	Activity (last month)	ρ	System Mass	Activity Momentum (last month)
MATHSE	19,255 (70,130)	0.0115	86.65	1,674,415 (6,076,765)
ESE	2,952 (13,751)	0.0344	29.07	85,815 (399,742)
BSE	246 (782)	0.0762	13.12	3,228 (10,260)
HSE	248 (1,110)	0.0554	18.10	4,489 (20,091)
15MW	1,999 (4,702)	0.0506	19.76	39,500 (92,912)
NLX	668 (1,131)	0.0532	18.80	12.558 (21,263)
NZ	12 (270)	0.0802	12.67	152 (3,421)
BP	54 (228)	0.0547	18.28	987 (4,168)

5. RELATED WORK

The work presented in this paper was inspired by and builds upon work presented in the areas of *critical mass theory* and *dynamical systems on networks*.

5.1. Critical Mass Theory

In 1985 and 1988, Oliver et al. [1985]; Oliver and Marwell [1988]; Marwell et al. [1988] have discussed and analyzed the concept of critical mass theory by introducing so called production functions to characterize decisions made by groups or small collectives. Fundamentally, these production functions represent the link between individual benefits and benefits for the group.

They argue that one very important aspect of critical mass is the natural limitation of collective goods for groups such as housing, food, fuel or oil. Hence, the capacity of users (and thus critical mass) for such a group or system is naturally limited by the corresponding resource. However, collective (digital) goods are not (or only artificially) limited for online communities; theoretically allowing for an infinite increase in users and interest. Without users motivated to contribute, interest will decrease and critical mass will lose momentum and ultimately decelerate until all interest vanishes. In their work they identified multiple different types of production functions, with the most important ones being: *Accelerating*, *decelerating* and *linear* functions. The idea behind accelerating production functions is that each contribution is worth more than its preceding one. In a decelerating production function the opposite would be the case, resulting in each succeeding contribution to be worth less than the preceding one, while contributions to linearly growing functions are always worth the same. Until today it is still mostly unclear what these production functions look like for online communities (e.g., StackOverflow) and online production systems (e.g., Semantic MediaWikis).

Depending on the investigated or desired point of view, different characteristics of these communities and online production systems can be used as basis for calculating production functions. The analysis of Oliver et al. [1985] also highlights that different production functions can lead to very different outcomes in similar situations. For example, given an accelerating production function, users who contribute to a system are likely to find their potential contribution "profitable", as each subsequent contribution increases the value of their own contribution. Naturally, this increases the incentive to make larger contributions to begin with. Given a deceleration production function, users would not immediately see the benefit of large contributions, given that each subsequent contribution is increasing the overall value less, while more effort, in the form of larger contributions, is needed to turn a decelerating production function into an accelerating one.

One approximation for critical mass by Solomon and Wash [2014] involved the investigation of the number of changes – as activity – and number of users – as growth of a community – for calculating production functions for WikiProjects. The authors argue that activity in online production systems, after certain amounts of time, is the best indicator of a self-sustaining system. In this work, we have extended the analysis presented by Solomon and Wash and specifically define the point of when an online system has reached critical mass and has become self-sustaining in terms of its activity dynamics. Walk and Strohmaier [2014] recently conducted a similar analysis to characterize critical mass for Semantic MediaWikis.

Raban et al. [2010] investigated factors that allow for a prediction of survival rates for IRC channels and identified the production function of these chat channels regarding the number of unique users versus the number of messages posted at certain times, as the best predictor.

Cheng and Bernstein [2014] have analyzed concepts of activation thresholds, which resemble features that, when achieved, can help to reach and sustain self-sustainability. They

created an online platform that allow groups to pitch ideas, which only will be activated if enough people commit to it.

With regards to activity, Suh et al. [2009] have shown that contributions to Wikipedia are slowing down, which is likely a direct consequence of the increase in required coordination activities, as well as comprehensive contribution guidelines which discourage posts by users. Kittur and Kraut [Kittur and Kraut 2008] have demonstrated that when reducing the overhead for editors—effectively minimizing the efforts necessary to contribute to Wikipedia—can help to increase the number of contributions and article quality. Similarly, Anderson et al. [2012] investigated the value and development of contributions to the question answering portal StackOverflow. In contrast, Yang et al. [2014] have investigated the evolution of two different types of users in StackOverflow, namely *sparrows* (very active users) and *owls* (experts) in the discussed topics, and could identify various differences between the two user-groups.

We use the notion of critical mass to define the barrier, that has to be overcome, for collaboration networks to become self-sustaining in terms of activity.

5.2. Dynamical Systems on Networks

Dynamical systems in a non-network context are a well-studied scientific and engineering field. Generally, a dynamical system is any system that changes in time, whose behavior is determined by some specific rules or (differential) equations over a set of quantifiable variables. We distinguish between continuous and discrete as well as deterministic and stochastic systems. Strogatz [1994] and Barrat et al. [2008] provide excellent introductions and analyses of dynamical systems.

Different social and economic processes, which take place both offline and online, have been modeled with the use of dynamical systems. In the context of the Web, the primary focus of dynamical systems was set on analyzing and understanding the diffusion of information in online social networks [Leskovec et al. 2007, 2009; Myers et al. 2012; Vespignani 2012], including the analysis of online memes and viral marketing.

On the other hand, the Bass Model [Bass 1976] describes how novel products are accepted and adopted in a network and has seen a wide variety of applications in different fields of research and also for practical use. The model consists of two parameters, the propensity for innovation and the propensity for imitation. A product will be successfully accepted and adopted by the community, depending in the ratio between these two parameters.

Acerbi et al. [2012] investigated factors that determine how social traits propagate within a specific popularity. Iribarren and Moro [2009] conducted a viral email experiment, allowing them to track the diffusion of information in a social network. They showed that due to heterogeneity in human activity, the most common and simple growth equation from epidemic models is not suitable to model information diffusion in social networks.

Recently, in the context of activity dynamics, Ribeiro [2014] conducted an analysis of the daily number of active users that visit specific websites, fitting a model that allows to predict if a website has reached self-sustainability, defined by the shape of the curve of the daily number of active users over time. He uses two constants α and β, where α represents the constant rate of active members influencing inactive members to become active. β describes the rate of an active member spontaneously becoming inactive. Whenever $\beta/\alpha \geq 1$ a website is unsustainable and without intervention the daily number of active users will converge to zero. If $\beta/\alpha < 1$ and the number of daily active users is initially higher than the asymptotic one, a website is categorized as self-sustaining.

The model presented in this paper to simulate activity dynamics heavily relies on the concept of dynamical systems on networks. We strongly believe that by modeling and understanding activity dynamics, we will gain a better understanding of the processes involved in and around the concept of peer influence in collaboration networks. Other areas of application for dynamical systems on networks are the modeling and simulation of diseases

in the form of *epidemic models*, and opinions or traits of a person, also known as *opinion dynamics*.

5.2.1. Epidemic Models. Modeling the outbreak of diseases can be seen as a special case of dynamical systems. At first, epidemic models dealt with the spreading of diseases in social (real life) networks [May and Anderson 1984; Hethcote 1978; Anderson and May 1991; Bolker and Grenfell 1993, 1995; Lloyd and May 1996; Keeling and Rohani 2002; Ferguson et al. 2003], ignoring the underlying network aspect, simulating contractions and outbreaks via random encounters of the whole population under investigation. For an exhaustive survey of epidemic models refer to Pastor-Satorras et al. [2014].

Henceforth, these models have been extended to include the structure and other aspects of the underlying networks [Rvachev and Longini 1985; Ferguson et al. 2003; Hufnagel et al. 2004; Longini et al. 2005; Ferguson et al. 2005; Colizza et al. 2006], limiting the spread and outbreaks according to different factors. Further, epidemic models were also utilized to simulate the spread for a plethora of properties in different kinds of networks, such as viruses spreading in computer networks [Kephart et al. 1993, 1997; Pastor-Satorras and Vespignani 2001b; Aron et al. 2002; Pastor-Satorras and Vespignani 2007] and information propagation (e.g., memes) [Leskovec et al. 2007] among others.

In general, epidemic models are based on the intuition that a disease propagates through a social network with a given infection rate, defining the probability that a neighbor of an already infected node contracts the disease. Different models have been developed and analyzed to simulate epidemic outbreaks in a population or network [Bailey et al. 1975; Anderson and May 1991; Hethcote 2000; Newman 2010], which can only transfer on contact. Typically, such an outbreak is modeled using a small number of possible states for each node and a fixed probability of contraction (e.g., β, γ), which defines the probability or "threshold" that has to be reached for a node to change to a different state. For example, the SI model consists of only two states – *susceptible* and *infected* – and one probability parameter β, that determines when the transition from susceptible to infected is initiated. Note that transitions in the SI model can only occur from susceptible to infected while already infected nodes remain infected indefinitely. As the infection rate is relative to the population under investigation, epidemic simulations with a small number of originally infected hosts usually start-off by slowly contracting the disease until exponential growth is reached. Once the majority of the population carries the disease, the infection process slows down again until the whole population is infected.

A more sophisticated extension to the SI model is the SIR model [Anderson and May 1991; Murray 2002], which additionally introduces the *recovered* (or *removed*) state as well as an additional parameter γ to model the transition from infected to recovered. Again, transitions only occur from susceptible to infected to recovered. As the name suggests, this newly introduced state allows nodes to become immune to the disease and will not be infected in the future, nor be able to infect other nodes. Other models for simulating epidemic outbreaks are the SIS and SIRS models, where the population can recover but does not become immune (SIS) or stays immune but still has a chance to become susceptible for infection again (SIRS) [Britton 2010; Dietz 1967].

Since their introduction, epidemic models have seen a wide array of application. For example, to analyze how computer viruses spread [Kephart and White 1991, 1993; Newman et al. 2002] or the study of epidemics in complex (scale-free, power-law) networks [Pastor-Satorras and Vespignani 2001b,a, 2002; Moreno et al. 2002].

Among others Wang et al. [2003] as well as Ganesh et al. [2005] demonstrated the importance of the networks spectra (eigenvalues and eigenvectors of the network adjacency matrix) for epidemic and dynamical network models [Chung et al. 2003a,b]. We show a similar dependency of activity dynamics on eigenvalues in this paper in Section 2.

5.2.2. Collective Behavior & Opinion Dynamics. Another important field of application of dynamical systems on networks are opinion dynamics. They are used to model collective behavior and influence, usually in the form of a consensus-reaching task, at every point in time. The main idea behind the concept of social influence is that interacting agents strive to become more alike [Festinger 1950].

For example, agents in the Ising model for ferromagnets [Binney et al. 1992; Barthélemy 2011] are influenced by the state/opinions of the majority of their peers. This influence naturally drives the system towards an ordered state where all agents are either positive or negative (ferromagnets). Hence, the model can be interpreted as a very simple model for simulating (binary) opinion dynamics. However, the transition probabilities of the Ising model are influenced by temperature, representing the modeling of external or influential factors. In particular, if the temperature is above a certain threshold, consensus-finding, in terms of magnetization, becomes an unstable process that never converges. The Potts model [Wu 1982; Dorogovtsev et al. 2008] further extends the Ising model by increasing the number of potential states an agent can assume from two (positive or negative) to an arbitrary number greater than two. Other factors that might influence the process of reaching consensus is the size of the system under investigation [Tessone and Toral 2009]. In particular, this means that differently sized (or connected) systems potentially need different strategies to reach consensus.

Opinions are usually represented as a set of words or numbers for each agent individually. Weidlich [1971] introduced such a model, based on sociodynamics, in 1971. Galam et al. [1982]; Galam and Moscovici [1991] analyzed the potential applications of the Ising model for simulating opinion dynamics starting in 1982.

The most wide-spread and adapted models to simulate (among others) opinion dynamics are the voter model [Clifford and Sudbury 1973; Holley and Liggett 1975], the Axelrod model [Axelrod 1997] as well as The Naming Game [Baronchelli et al. 2006].

The voter model constitutes that each agent is equipped with a binary variable. At each step in time, the binary variable of one (randomly chosen) agent is synchronized with one of its neighbors variable. Introducing the concept of social influence for opinion dynamics. The voter model has since been adapted and extended by many researchers to fit an array of different purposes (e.g., [Mobilia 2003; Mobilia and Georgiev 2005; Mobilia et al. 2007; Vazquez et al. 2003; Vazquez and Redner 2004; Castelló et al. 2006]).

The Axelrod model [Axelrod 1997] combines the notion of social influence – individuals becoming more similar upon frequent interactions – and the tendency that similar individuals will have a higher tendency (and frequency) to interact with each other. Each agent is endowed with a set of characterizing variables. The more variables are shared among two agents, the more similar they are. Given this description, one would assume that the described notions are self-reinforcing dynamics and hence, will inevitably produce stable networks with only identical agents. However, Castellano et al. [2000] have shown that the resulting number of different states is dependent on the number of characterizing variables. Large numbers are likely to result in very few similar individuals (high agent diversity). Analogously to the voter model, the Axelrod model has been extensively adapted, analyzed and expanded by researchers to broaden our understanding of the spread of (cultural) traits across agents (e.g., Klemm et al. [2003b,a]; Flache and Macy [2007]).

The Naming Game originates from idea to analyze and explore the evolution of language [Steels 1995]. Baronchelli et al. [2006] introduced the most basic version of The Naming Game in 2006, where a group of agents that communicate via a complete network, try to reach consensus when naming an entity. Each agent holds a list of synonyms or words associated with the entity, also referred to as vocabulary, under investigation. Every iteration (or step in time), two agents are chosen. One agent is assigned the role of the speaker, who randomly choses a word of a given/pre-defined vocabulary. If the other agent – the listener – knows (i.e., also has the word in the vocabulary) the chosen word, both agents discard

all other words in their vocabulary and "agree" on the common word. However, if the listeners do not know the word of the speaker, the word is appended to their vocabulary and no words are discarded. In the next step another pair of nodes is chosen and process is repeated until either consensus is found or a predetermined number of steps (time) have passed. The Naming Game has spurred a complete line of dynamical models with a variety of different parameters, that each address different problems and tasks (e.g., Abrams and Strogatz [2003]; Minett and Wang [2008]; Wang and Minett [2005]; Castelló et al. [2006]). For an excellent and comprehensive introduction to opinion dynamics (among others) we refer the interested reader to Castellano et al. [2009].

6. DISCUSSION, LIMITATIONS & FUTURE WORK

We have developed a model[12] to simulate and characterize the intricate dynamics of activity in collaboration networks, consisting of an *Activity Decay Rate* and *Peer Influence Growth Rate*. First, we applied it on Zachary's Karate Club (see Figure 3) dataset to illustrate its core mechanics. Subsequently, we continued with a linear stability analysis (cf. Section 2.2) and depicted the behavior that can occur when the master stability equation is invalidated (see Figure 3). Using our proposed model to simulate activity dynamics, we have shown that the overall activity in collaboration networks appears to be a composite of the *Activity Decay Rate* and the *Peer Influence Growth Rate*, as described in Section 2. In Section 3, we have fitted our model on synthetic and empirical datasets to simulate activity dynamics trends.

The presented results are destined to be interpreted only and solely as an indicator for trends in activity dynamics, rather than absolute values that can be used for accurately predicting the activity for a given system. This is a direct result of the different approximations and simplifications (cf. Section 3) that we have made when estimating the parameters for our activity dynamics simulation.

Note that one advantage of our model over other existing approaches, such as autoregression, is the interpretability of the ratio λ/μ. For example, a ratio of 4 means that users intrinsically lose activity 4 times faster than they can get back from one of their peers, while the coefficients of the autoregression lack such interpretable characteristics. Further, using the concept of dynamical systems we can represent the underlying mechanisms in a closed form, allowing for detailed analytical analyses (i.e., the linear stability analysis), which is much harder (if not impossible) to conduct for other models, such as agent-based models, autoregression or more complex models based on dynamical systems.

For future work we plan on extending the ability of our model to not only reflect on changes in activity dynamics but also properly cope with structural changes in the underlying collaboration networks. One additional limitation of the presented approach is the fact that nodes with a very small degree, which are not connected to the largest connected component, inevitably will lose activity until they reach the point of total inactivity. Including the structural evolution of a collaboration network in our analyses will allow us to mitigate this effect, as users will only be added to the collaboration network and considered in our calculations, once they have actually become active. One potential approach involves the investigation of snapshots of the collaboration networks at every τ, providing additional insights into the evolution of the parameters of our model and the investigated systems. Additionally, we assume that peer influence is a symmetric property. This means that posts and replies exercise the same amount of influence on peers as we do not differentiate between different types of activity and influence will always traverse along both directions of the edges in our collaboration networks. Further, tasks that do not trigger entries in the

[12]We have released a Python implementation of our model, to estimate empirical parameters and run activity dynamics simulations, as Open Source Software at https://github.com/simonwalk/ActivityDynamics.

change-logs (i.e., reading articles, posts or replies) are not considered in our experiments due to a lack of available data.

The fact that the *Activity Dynamics Model* only requires a single parameter to be configured represents not only an advantage, but also a limitation. Given that there is only one parameter that determines the evolution of activity in a system, we are not be able to model periodic fluctuations with only one ratio. Instead, we have to calculate ratios for multiple points in time. For future work we plan on extending the *Activity Dynamics Model* by adding parameters, for example, to model different external influences. With this extended model, we will be able to simulate such periodic patterns with a single configuration. On the other hand, we are only able to model additional (social) mechanisms with the use of additional parameters. For example, one reason for the decreasing levels of activity in Wikipedia might also be related to a very high barrier for newly registered users to add content due to comprehensive guidelines for contributions and a very concentrated and active community of power users. Over time, these power users leave Wikipedia for various reasons while new contributors are lacking to fill in the gaps.

Furthermore, all of our estimated parameters are calculated for the collaboration networks as a whole. Future work will also include extending the activity dynamics model to calculate the ratio λ/μ on a user level, rather than on a network level. This modification not only potentially increases the accuracy of our model but would also allow us to gather additional information for each user of the corresponding networks. Further, with an increased accuracy in our simulations it will be possible to conduct activity prediction experiments and emulate network attacks as well as optimize (arbitrary) cost-strategies for increasing activity in these systems.

In this context it is also worth mentioning that decreasing levels of activity for collaboration networks can also signal that the community has completed their work and no further actions are required as the intended goal has been achieved. Further analyses are required to determine if completeness and quality of content affect activity in collaboration networks. One could even argue that, once we are able to calculate λ/μ for each user, we could potentially observe the evolution of users and categorize different types of users in collaboration networks (e.g., early adopters or experienced users versus new and inexperienced users).

The ratio λ/μ—describing how fast users lose activity (*Activity Decay Rate* λ) over how fast they regains activity over their neighbors (*Peer Influence Growth Rate* μ)—fluctuates below the corresponding highest eigenvalue κ_1 for all investigated empirical datasets. Negative peaks in this ratio represent periods of time (τ; in our case weeks) where activity grew faster than could be compensated by the *Peer Influence Growth Rate*. It naturally follows that a decrease of λ—resulting in less activity-loss per contribution for each user—is necessary to accomplish such drastic increases of activity. If the network itself is of a smaller scale and/or these negative peaks occur on a frequent basis, the activity dynamics of the corresponding networks are depending on the contributions (and thus influence) of single (individual) users. To compare the stability of the activity dynamics across multiple networks we calculated the *System Mass* and *Activity Momentum* p—indicating the required force to accelerate or render the corresponding collaboration networks inactive.

When comparing p and the results of our empirical illustration (cf. Figures 9 and 10) between the different datasets, we can see that the *Activity Momentum* is very small for datasets that either (i) exhibit only a very small number of changes and are close to inactivity or (ii) exhibit a small κ_1 (see Figure 9 and 10). This suggests that we can use *Activity Momentum* as an indicator for the robustness of a collaboration network with regards to its activity dynamics.

Further, we can characterize the potential of a collaboration network to become self-sustaining by comparing the calculated ratios of λ/μ with the corresponding κ_1 and *Activity Momentum*. If the ratio is below κ_1, our master stability equation is invalidated, pushing the system towards a new fixed point where the forces of the *Activity Decay Rate* and the *Peer*

Influence Growth Rate reach an equilibrium so that the network converges towards a state of immanent and lasting activity (see Figure 3). If such a state is reached and combined with a high *Activity Momentum*, the corresponding collaboration network has reached critical mass of activity and has become self-sustaining; no external impulses are required to keep the network active. Of course, in real world scenarios, activity will not last forever without providing additional incentives as interest (and thus activity) in a system potentially decays over time. As a consequence, this would first result in an increase of μ and inevitably, with a sufficiently large μ, the collaboration network would return to its stable fixed point, once our master stability equation holds again, and activity would once more converge towards zero. Once we extend our model to allow for user-based calculations, we will be able to not only calculate *Activity Momentum* for collaboration networks, but also for single and individual users.

ACKNOWLEDGMENTS

This research was in part funded by the FWF Austrian Science Fund research projects P24866.

REFERENCES

Daniel M Abrams and Steven H Strogatz. 2003. Linguistics: Modelling the dynamics of language death. *Nature* 424, 6951 (2003), 900–900.

Alberto Acerbi, Stefano Ghirlanda, and Magnus Enquist. 2012. The logic of fashion cycles. *PloS one* 7, 3 (2012), e32541.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 850–858.

R. M. Anderson and R. M. May. 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, USA.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.

Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341.

Joan L Aron, Michael O'leary, Ronald A Gove, Shiva Azadegan, and M.Cristina Schneider. 2002. The Benefits of a Notification Process in Addressing the Worsening Computer Virus Problem: Results of a Survey and a Simulation Model. *Comput. Secur.* 21, 2 (March 2002), 142–163. DOI:http://dx.doi.org/10.1016/S0167-4048(02)00210-9

Robert Axelrod. 1997. Advancing the art of simulation in the social sciences. In *Simulating social phenomena*. Springer, 21–40.

Norman TJ Bailey and others. 1975. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications* 311, 3 (2002), 590–614.

Andrea Baronchelli, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels. 2006. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment* 2006, 06 (2006), P06014.

Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. 2008. *Dynamical processes on complex networks*. Vol. 1.

Marc Barthélemy. 2011. Spatial networks. *Physics Reports* 499, 1 (2011), 1–101.

Frank M Bass. 1976. A New Product Growth Model for Consumer Durables. In *Mathematical Models in Marketing*. Springer, 351–353.

James J Binney, NJ Dowrick, AJ Fisher, and M Newman. 1992. *The theory of critical phenomena: an introduction to the renormalization group*. Oxford University Press, Inc.

BM Bolker and BT Grenfell. 1993. Chaos and biological complexity in measles dynamics. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 251, 1330 (1993), 75–81.

Benjamin Bolker and Bryan Grenfell. 1995. Space, persistence and dynamics of measles epidemics. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 348, 1325 (1995), 309–320.

Tom Britton. 2010. Stochastic epidemic models: a survey. *Mathematical biosciences* 225, 1 (2010), 24–35.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81, 2 (2009), 591.

Claudio Castellano, Matteo Marsili, and Alessandro Vespignani. 2000. Nonequilibrium phase transition in a model for social influence. *Physical Review Letters* 85, 16 (2000), 3536.

Xavier Castelló, Víctor M Eguíluz, and Maxi San Miguel. 2006. Ordering dynamics with two non-excluding options: bilingualism in language competition. *New Journal of Physics* 8, 12 (2006), 308.

Justin Cheng and Michael S Bernstein. 2014. Catalyst: Triggering Collective Action with Thresholds. (2014).

Nicholas A Christakis and James H Fowler. 2008. The collective dynamics of smoking in a large social network. *New England journal of medicine* 358, 21 (2008), 2249–2258.

Fan Chung, Linyuan Lu, and Van Vu. 2003a. Eigenvalues of random power law graphs. *Annals of Combinatorics* 7, 1 (2003), 21–33.

Fan Chung, Linyuan Lu, and Van Vu. 2003b. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* 100, 11 (2003), 6313–6318.

Peter Clifford and Aidan Sudbury. 1973. A model for spatial conflict. *Biometrika* 60, 3 (1973), 581–588.

Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7 (2006), 2015–2020.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 307–318. http://dl.acm.org/citation.cfm?id=2488388.2488416

Klaus Dietz. 1967. Epidemics and rumours: A survey. *Journal of the Royal Statistical Society. Series A (General)* (1967), 505–528.

Sergey N Dorogovtsev, Alexander V Goltsev, and José FF Mendes. 2008. Critical phenomena in complex networks. *Reviews of Modern Physics* 80, 4 (2008), 1275.

Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S Burke. 2005. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437, 7056 (2005), 209–214.

Neil M Ferguson, Matt J Keeling, W John Edmunds, Raymond Gani, Bryan T Grenfell, Roy M Anderson, and Steve Leach. 2003. Planning for smallpox outbreaks. *Nature* 425, 6959 (2003), 681–685.

Leon Festinger. 1950. *Social pressures in informal groups: A study of human factors in housing.* Stanford University Press.

Andreas Flache and Michael W Macy. 2007. Local Convergence and Global Diversity: The Robustness of Cultural Homophily. *arXiv preprint physics/0701333* (2007).

Serge Galam, Yuval Gefen, and Yonathan Shapir. 1982. Sociophysics: A new approach of sociological collective behaviour. I. mean-behaviour description of a strike. *Journal of Mathematical Sociology* 9, 1 (1982), 1–13.

Serge Galam and Serge Moscovici. 1991. Towards a theory of collective phenomena: consensus and attitude changes in groups. *European Journal of Social Psychology* 21, 1 (1991), 49–74.

Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. 2005. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, Vol. 2. IEEE, 1455–1466.

Herbert W Hethcote. 1978. An immunization model for a heterogeneous population. *Theoretical population biology* 14, 3 (1978), 338–349.

Herbert W Hethcote. 2000. The mathematics of infectious diseases. *SIAM review* 42, 4 (2000), 599–653.

Richard A Holley and Thomas M Liggett. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability* (1975), 643–663.

Lars Hufnagel, Dirk Brockmann, and Theo Geisel. 2004. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America* 101, 42 (2004),

15124–15129.
José Luis Iribarren and Esteban Moro. 2009. Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters* 103, 3 (2009), 038702.
Matt J Keeling and Pejman Rohani. 2002. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters* 5, 1 (2002), 20–29.
Jeffrey O Kephart, Gregory B Sorkin, David M Chess, and Steve R White. 1997. Fighting Computer Viruses: Biological Metaphors Offer Insights into Many Aspects of Computer Viruses and Can Inspire Defenses Against Them. *Scientific American* (1997).
Jeffrey O Kephart and Steve R White. 1991. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*. IEEE, 343–359.
Jeffrey O Kephart and Steve R White. 1993. Measuring and modeling computer virus prevalence. In *Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on*. IEEE, 2–15.
Jeffrey O Kephart, Steve R White, and David M Chess. 1993. Computers and epidemiology. *Spectrum, IEEE* 30, 5 (1993), 20–26.
Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 37–46.
Konstantin Klemm, Víctor M Eguíluz, Raúl Toral, and Maxi San Miguel. 2003a. Global culture: A noise-induced transition in finite systems. *Physical Review E* 67, 4 (2003), 045101.
Konstantin Klemm, Víctor M Eguíluz, Raúl Toral, and Maxi San Miguel. 2003b. Nonequilibrium transitions in complex networks: A model of social interaction. *Physical Review E* 67, 2 (2003), 026120.
Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. 2007. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)* 1, 1 (2007), 5.
Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 497–506.
CC Lin and LA Segel. 1988. *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Vol. 1. SIAM.
Alun L Lloyd and Robert M May. 1996. Spatial heterogeneity in epidemic models. *Journal of theoretical biology* 179, 1 (1996), 1–11.
Ira M Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek AT Cummings, and M Elizabeth Halloran. 2005. Containing pandemic influenza at the source. *Science* 309, 5737 (2005), 1083–1087.
Gerald Marwell, Pamela E Oliver, and Ralph Prahl. 1988. Social Networks and Collective Action: A Theory of the Critical Mass, Ill. *Amer. J. Sociology* 94, 3 (1988), 502–534.
Robert M May and Roy M Anderson. 1984. Spatial heterogeneity and the design of immunization programs. *Mathematical Biosciences* 72, 1 (1984), 83–111.
James W Minett and William SY Wang. 2008. Modelling endangered languages: The effects of bilingualism and social structure. *Lingua* 118, 1 (2008), 19–45.
Mauro Mobilia. 2003. Does a single zealot affect an infinite group of voters? *Physical Review Letters* 91, 2 (2003), 028701.
Mauro Mobilia and Ivan T Georgiev. 2005. Voting and catalytic processes with inhomogeneities. *Physical Review E* 71, 4 (2005), 046102.
Mauro Mobilia, A Petersen, and Sidney Redner. 2007. On the role of zealotry in the voter model. *Journal of Statistical Mechanics: Theory and Experiment* 2007, 08 (2007), P08029.
Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2002. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems* 26, 4 (2002), 521–529.
J D Murray. 2002. *Mathematical biology*. Springer, New York. http://www.worldcat.org/search?qt=worldcat_org_all&q=9780387952239
Seth A Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 33–41.
Mark Newman. 2010. *Networks: an introduction*. Oxford University Press.
Mark EJ Newman. 2001. Scientific collaboration networks. I. Network construction and fundamental results.

Physical review E 64, 1 (2001), 016131.

Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. 2002. Email networks and the spread of computer viruses. *Physical Review E* 66, 3 (2002), 035101.

Pamela Oliver, Gerald Marwell, and Ruy Teixeira. 1985. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American journal of Sociology* (1985), 522–556.

Pamela E Oliver and Gerald Marwell. 1988. The Paradox of Group Size in Collective Action: A Theory of the Critical Mass. II. *American Sociological Review* (1988), 1–8.

Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2014. Epidemic processes in complex networks. *arXiv preprint arXiv:1408.2701* (2014).

Romualdo Pastor-Satorras and Alessandro Vespignani. 2001a. Epidemic dynamics and endemic states in complex networks. *Physical Review E* 63, 6 (2001), 066117.

Romualdo Pastor-Satorras and Alessandro Vespignani. 2001b. Epidemic spreading in scale-free networks. *Physical review letters* 86, 14 (2001), 3200.

Romualdo Pastor-Satorras and Alessandro Vespignani. 2002. Epidemic dynamics in finite size scale-free networks. *Physical Review E* 65, 3 (2002), 035108.

Romualdo Pastor-Satorras and Alessandro Vespignani. 2007. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press.

Daphne R. Raban, Mihai Moldovan, and Quentin Jones. 2010. An Empirical Study of Critical Mass and Online Community Survival. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 71–80. DOI:http://dx.doi.org/10.1145/1718918.1718932

Bruno Ribeiro. 2014. Modeling and Predicting the Growth and Death of Membership-based Websites. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 653–664. DOI:http://dx.doi.org/10.1145/2566486.2567984

Leonid A Rvachev and Ira M Longini. 1985. A mathematical model for the global spread of influenza. *Mathematical biosciences* 75, 1 (1985), 3–22.

Jacob Solomon and Rick Wash. 2014. Critical Mass of What? Exploring Community Growth in WikiProjects. (2014).

Luc Steels. 1995. A self-organizing spatial vocabulary. *Artificial life* 2, 3 (1995), 319–332.

Steven H. Strogatz. 1994. *Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering (Studies in Nonlinearity)*. Perseus Books Group.

Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 8.

Claudio J Tessone and Raul Toral. 2009. Diversity-induced resonance in a model for opinion formation. *The European Physical Journal B-Condensed Matter and Complex Systems* 71, 4 (2009), 549–555.

Federico Vazquez, Paul L Krapivsky, and Sidney Redner. 2003. Constrained opinion dynamics: Freezing and slow evolution. *Journal of Physics A: Mathematical and General* 36, 3 (2003), L61.

F Vazquez and S Redner. 2004. Ultimate fate of constrained voters. *Journal of Physics A: Mathematical and General* 37, 35 (2004), 8479.

Alessandro Vespignani. 2012. Modelling dynamical processes in complex socio-technical systems. *Nature Physics* 8, 1 (2012), 32–39.

Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)* (2012), 2.

Simon Walk and Markus Strohmaier. 2014. Characterizing and Predicting Activity in Semantic MediaWiki Communities. In *SWCS14 Third International Workshop on Semantic Web Collaborative Spaces, 2014*. 21.

William SY Wang and James W Minett. 2005. The invasion of language: emergence, change and death. *Trends in ecology & evolution* 20, 5 (2005), 263–269.

Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. 2003. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*. IEEE, 25–34.

Wolfgang Weidlich. 1971. THE STATISTICAL DESCRIPTION OF POLARIZATION PHENOMENA IN SOCIETY. *Brit. J. Math. Statist. Psych.* 24, 2 (1971), 251–266.

Fa-Yueh Wu. 1982. The potts model. *Reviews of modern physics* 54, 1 (1982), 235.

Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. 2014. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *User Modeling, Adaptation, and Personalization*. Springer, 266–277.