Can Epidemic Models Describe the Spread of Research Topics Across Disciplines?

Istvan Z. Kiss[1], Mark Broom[1] and Ismael Rafols[2]

[1] *i.z.kiss@sussex.ac.uk, m.broom@sussex.ac.uk*
Department of Mathematics, University of Sussex, Brighton, Falmer, BN1 9RF (England)

[2] *i.rafols@sussex.ac.uk (*corresponding author)
SPRU –Science and Technology Policy Research, University of Sussex, Brighton, BN1 9QE (England)

Abstract

This paper introduces a new approach to describe the spread of research topics across disciplines using epidemic models. The approach is based on applying individual-based models from mathematical epidemiology to the diffusion of a research topic over a contact network that represents the map of science –as obtained from citations between ISI Subject Categories. Using research publications on *kinesin* as a case study, we report a better fit between model and empirical data when using the citation-based contact network. Incubation periods on the order of 4 to 15.5 years support the view that, whilst research topics may grow very quickly, they face difficulties to overcome disciplinary boundaries.

Introduction

The diffusion of knowledge in general and the spread of research topics in particular, resemble in many ways the spread of infectious diseases where contact between an infectious and a susceptible individual can lead to the spread of infection. In a similar way, individuals or scientific communities working on a particular research topic or idea can motivate other individuals or groups to start work based on the same or similar topics. Using models that are well known in the context of mathematical epidemiology, we develop an individual-based weighted network model that captures the dynamics of transmission of a research topic across scientific disciplines. The novelty of the approach we present here is that, whereas previous studies have investigated the growth of a topic in terms of networks of papers (e.g. Bettencourt et al., 2006; 2008), we inquire here into how a research topic spreads over disciplines. This novel perspective captures the diffusion of topics over the map of science, whereas former studies had focused on growth dynamics.

In this exploratory study, we show that the spread of a topic (in this case, research on molecular motor *kinesin* –a "nano-engine") over a network of disciplines can be well approximated by models used in the context of the transmission of infectious diseases (Keeling & Rohani, 2008). Our approach builds on previous network models that have been successfully used to explain and predict the pattern of infectious disease transmission (Kiss et al., 2006; Green et al., 2008).

Methods and data

A set of publications (articles, reviews and letters) related to the molecular motor kinesin was constructed searching the term "kinesin" in the bibliographic field "Topic" of the ISI Web of Science database. This search yielded 5,260 publications starting from 1985 (2 publications) to 2007 (557)[i]. Each publication is assigned to a discipline according to ISI Web of Science's classification in terms of Subject Categories (SCs). The matrix of citations between SCs was obtained from Leydesdorff and Rafols' (2009) (data available at Leydesdorff's webpage)[ii]. This SC to SC citation matrix had been created for 2006 from the Journal Citation Reports (JRC) of the Science Citation Index (SCI).

Proceedings of ISSI 2009, pages 857–861.
Edited by B. Larsen and J. Leta.

This citations matrix with 171 nodes (N=171) as SCs[iii] is used to construct the contact network over which the transmission dynamics unfolds (in this case, the spread of research topics). The baseline citation network (the "backbone" of science) may be understood as representing the knowledge flows among SCs. The key assumption in the model is that that the weight of a link of this network is a good indicator in determining the likelihood of a SC becoming research-active in a certain area given that some other related SCs are already research-active in this specific area. However, the links are normalised so that the weight of the incoming links for all SCs add up to one. Hence the weight w_{ij} of the directed link from SC_i to SC_j is:

$$w_{ij} = \frac{\#\text{ of citations from SC}_j\text{ to SC}_i}{\#\text{ of total citations made by SC}_j} \quad \text{with} \quad \sum_i w_{ij} = 1 \text{ for } \forall j = 1 \ldots N. \quad (1)$$

The model

Although the quantitative modelling of the spread of research ideas or topics using epidemiological models started almost as early as bibliometrics (see review by Garfield. 1980), it has been recently re-opened with more sophisticated methods (Bettencourt et *al.* 2006; 2008). Here we suggest a perspective that departs from these studies in two respects. First, we aim to investigate the spread of a topic over disciplines (SCs) –not publications or authors, which describe mainly just growth. Second, the former studies have used only simple differential-equation-based compartmental models. While compartmental models are transparent and allow the derivation of some analytical results, they are limited in their capability to capture heterogeneities at the individual level (i.e. in this case heterogeneity in the in and out degree distributions and link weight distribution).

In our model, we look if one SC (a node in the network) has publications in a topic at a given time, and explore how this spreads to other nodes (or SCs) in the map of science via links weighted according to intensity of citations between SCs. Following Sharkey (2008), we use an individual-based model where equations for the probability of being in a particular state at a particular time are worked out based on the links between SCs, the status of neighbours, and given transmission and incubation rates.

In the models that we consider, any SC is in one of the three possible states: susceptible to infection (S), incubating the topic (E), and finally infected or adopter (I). Susceptible SCs are either not aware of a particular research topic or if aware, can still choose not to adopt it. Incubating SCs are those that are aware of a certain topic and may have started engaging with it. This is expected to result in tangible research output in the form of papers. Infected SCs or adopters are those that are actively working and publishing in a particular research topic. In terms of empirical data, infected SCs in a given year are those that published at least one paper on kinesin during that year. Further states such as recovered (i.e. SCs that have stopped working on a particular research area, often denoted by R in some models) and sceptics or stiflers (i.e., SCs that are aware of the topic but do not engage with it or support another competing topic or research area, often denoted by Z) are possible. However, in our current model the recovered class is ignored since the empirical data for our case shows continued growth (i.e. all SCs continue to publish in the topic once they have become active). The models are called SEI when they include all three states or SI, when only S and I states are included. The model equations are given by:

$$\begin{cases} dP_{S(i)}(t)/dt = -\sum_j T_{ji} P_{I(j)}(t) P_{S(i)}(t), \\ dP_{E(i)}(t)/dt = \sum_j T_{ji} P_{I(j)}(t) P_{S(i)}(t) - g P_{E(i)}(t), \\ dP_{I(i)}(t)/dt = g P_{E(i)}(t), \end{cases} \quad (2)$$

where $0 \leq P_{I(j)}(t) \leq 1$ denotes the probability of node j being infected at time t (likewise for $E(j)$ and $S(j)$). Throughout the simulation, $P_{S(j)}(t)+P_{E(j)}(t)+P_{I(j)}(t)=1$, for all $\forall t>0$. The contact network is represented by $T_{ji} = \tau G_{ji}$ with $G_{ji}=(w_{ji})_{j,i=1,\ldots,N}$ denoting the adjacency matrix that includes link weights. τ is the transmission rate per contact and $1/g$ is the average incubation or latent period. By numerically integrating the ordinary differential equations, the number of the infected or adopter SCs at time t according to the model can be estimated as $I(t) = \sum_j P_{I(j)}(t)$.

The simulation is started at time $t = 0$ corresponding to 1985 and the equations are integrated forward in time until 2007. The initial infection is seeded in the two SCs corresponding to *Cell Biology*, and *Biochemistry and Molecular Biology*. The SEI model has two free parameters (i.e. τ and g) that allow to fit the model output to the empirical data. In this case SC count and $I(t)$ are normalised by N and compared on a yearly basis. The estimation of parameters is performed according to a modified version of the Kolmogorov-Smirnov statistics, i.e. a minimum distance estimation between an empirical distribution function of a sample and the cumulative distribution function of the reference distribution

$$AdaptedKS = \sup_{t=1985,1986,\ldots,2007} \left[\frac{1}{N} \left(\sum_{j=1}^{N} P_{I(j)}(t) - EmpiricalCount(t) \right) \right]. \quad (3)$$

The count of the SCs that are active in kinesin-related research provides information at the level of all SCs or population level. However, a good model will also be able to provide information at the individual, or in this case, SC level. An appropriate model that fits the data well, apart from accurate prediction of the growth in the number of SCs, can also predict the exact SCs that are active, at a particular time, in the kinesin-related research. To monitor model prediction at the SC level the following likelihood function is considered:

$$L = \frac{1}{M} \sum_{t=1985,1986,\ldots,2007} \left(-\log \left(\prod_{i=1}^{N} \left(P_{I(i)}(t) \right)^{Y_i(t)} \left(1 - P_{I(i)}(t) \right)^{1-Y_i(t)} \right) \right), \quad (4)$$

where $Y_i(t) \in \{0,1\}$, with a value of one denoting a SC that is active in kinesin-related research at time t and zero if otherwise (Green et al., 2008). M denotes the number of time points where comparisons at the individual level are made. In this case $M=23$, and this corresponds to yearly comparisons from 1985 until 2007.

Results

To explore the significance of the weighted contact network (the "backbone" of science) and the incubation time in explaining the spread of research topics, we investigate different models, as shown in Table 1. We consider three different versions of the SI model without the incubation time g. First, the case where all links are equal to the average link weight. Second, the case where the weights of all incoming links of any node or SC are equal and sum to one. Third, the case in which the network is empirically weighted. Finally, we consider the full SEI model, as shown in Equation 2.

Table 1 shows that the SEI model with the empirically weighted network (i.e. using the backbone of science to model "flow of knowledge") provides a better fit (lower *AdaptedKS* and *L)* compared to the case when all weights are assumed to be equal. Figure 1 illustrates the best fit prevalence curves based on *AdaptedKS* (population level) and *L (*individual SC level).

Table 1: Parameter estimates for different network and disease transmission models.

Model type	Network Type	*Model parameters*		*Descriptors of model fit*	
		τ	*1/g*	***AdaptedKS*** *(Populat. level)*	***L*** *(Individual level)*
SI	All weights equal	0.229	---	Min=0.1146	1700.530081
		0.238	---	0.1533	Min=1694.533715
SI	Weights of all incoming links sum to one	0.230	---	Min=0.1153	1811.015665
		0.238	---	0.1484	Min=1806.257798
SI	Empirically weighted	0.174	---	Min=0.0518	Min=1395.716852
SEI	Empirically weighted	0.37	4.0	0.0872	**Min=1358.911193**
		1.90	15.5	**Min=0.0261**	1460.664702

Note: The best fit model and optimal parameters are indicated in bold type.

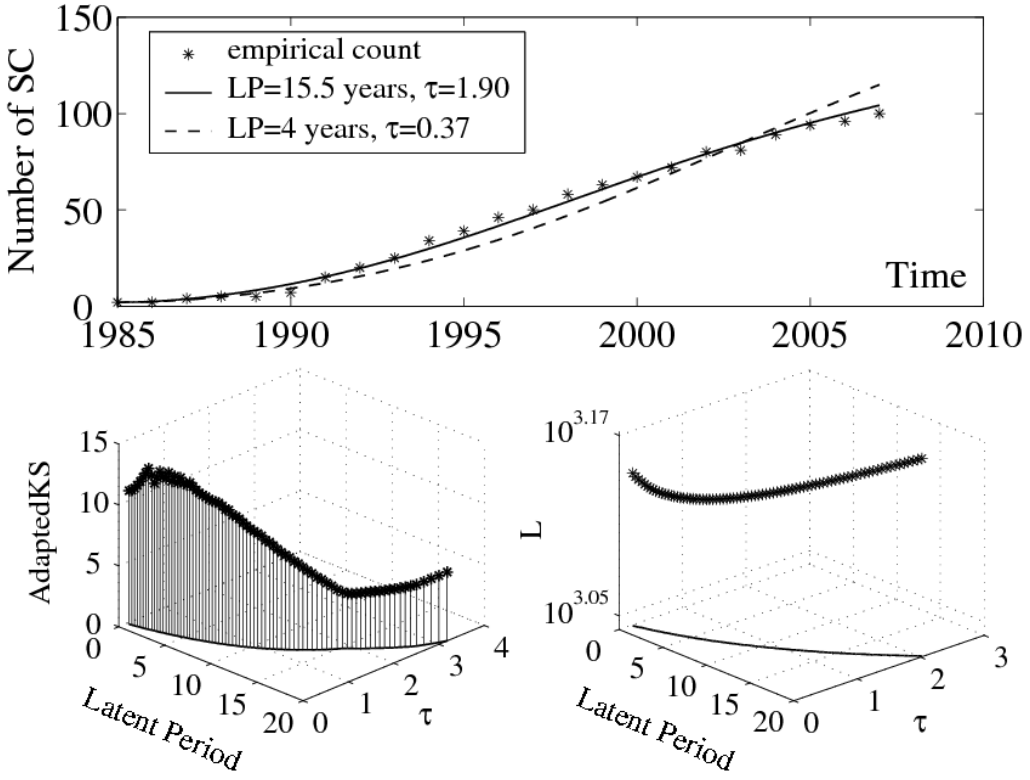

Figure 1: Best fit curves to the growth of the number of SCs that are active in kinesin-related research (top), *AdaptedSK* (bottom left) and *L* (bottom right), as a function of the latent period (1/*g*) and transmission rate (τ). Model based on the weighted network with SEI transmission.

To interpret the results shown in Figure 1, it is useful to consider first a fixed latent period 1/*g* and second the value of τ that minimises the difference between data and model output. Long latency periods require high values of the transmission rate *τ,* since individuals remain exposed without becoming infectious. This tendency is reflected by a set of optimal parameter pairs (1/*g, τ*) with both latent period and transmission rate increasing simultaneously (Figure

1, bottom left and right). However, the goodness of fit along the set of optimal pairs changes with the best agreement based on *AdaptedKS* occurring for $(1/g, \tau) = (15.5, 1.90)$. For longer latent periods, this measure indicates that the discrepancy between model output and data increases (Figure 1, bottom left). A similar tendency is observed when the parameter estimation happens based on *L* with the best agreement between data and model output for $(1/g, \tau) = (4.0, 0.37)$. Table 1 shows that while the agreement at the population level (*AdaptedKS*) is much better for the SEI model, for the same pair of parameters the agreement at the individual level (*L*) is not as good as for the simple *SI* model. Hence, agreement at both individual and population level is difficult to obtain.

Discussion

This paper has demonstrated the feasibility of applying individual-based epidemiological models to the spread of a research topic over disciplines. Using research on kinesin as a case study, we have confirmed that agreement between model output and empirical data significantly increases when the normalised weighted contact network between SCs is used (the backbone of science). We have found incubation periods on the order of 4 to 15.5 years, which support the view that, whilst research topics may grow very quickly, they face difficulties to overcome disciplinary boundaries. This type of information regarding the diffusion rate of research topics over disciplines may be of particular interest for emergent fields, such as nanotechnologies or biotechnologies, that are viewed as not conforming to traditional bodies of knowledge. However, the model agreement at individual level can be further improved by considering internal SC dynamics (since most topic growth happens within some SCs, this strongly dampens diffusion), and by taking into account not only whether a SC is infected or not, but also how active it is (as shown by number of publications in the topic).

References

Garfield, E. (1980). The epidemiology of knowledge and the spread of scientific ideas. *Current Contents*, 35, 5–10.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C. & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3), 495-518.

Bettencourt, L.M.A., Cinrón-Arias, A., Kaiser, D.I. & Castillo-Chávez, C. (2006). The power of a good idea : Quantitative modeling of the spread of ideas from epidemiological models, *Physica A* (2006) 364, 513-536.

Green, D.M., Kiss, I.Z., Mitchell, A.P. & Kao, R.R. (2008). Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings of the Royal Society B,* 275, 1001-1005.

Keeling, M.J. & Rohani, P. (2008). *Modelling infectious disease in humans and animals*, Princeton, NJ: Princeton University Press.

Kiss, I.Z. Green D.M. & Kao, R.R. (2006). The network of sheep movements within Great Britain: network properties and their implication for infectious disease spread. *Journal of the Royal Society Interface*, 3, 669-677.

Leydesdorff, L. & Rafols, I. (2009) A global map of science based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology*. 60(2), 348-362.

Sharkey, K. J. (2008). Deterministic epidemiological models at the individual level, *Journal of Mathematical Biology*, 57, 311-331.

[i] Due to improved indexing since 1991, this search underestimates the number of publications until 1990 –an effect we will overlook here.

[ii] Data available at http://www.leydesdorff.net/map06/data.xls.

[iii] The SCI had 172 SCs in 2006. We removed the SC "Multidisciplinary Sciences" because we understood that it might lead to misleading linkages, given that the publication patterns of journals such as *Nature* or *Science* publish for diverse audiences but do not necessarily *connect* them.