Knowledge epidemics

Nikolay K. Vitanov and Marcel Ausloos

Plan of presentation

- Knowledge, qualitative research, historical remarks, kinds of models
- Large size systems: deterministic models
- Small size systems: stochastic models
- Statistical approach: distributions of Lotka, Pareto, Bradford and Zipf

Any diffusion of ideas is closely connected to the creation and diffusion of knowledge and to the technological evolution of society. Thus knowledge creation and exchange and its transformation to innovations and economic growth, in recent times, is first described from a historical point of view. Next, three approaches are discussed for modelling the diffusion of ideas in the areas of science and technology, so called (i) deterministic, (ii) stochastic, and (iii) statistical approaches. They are illustrated through corresponding numerous population dynamics and epidemic models relative to the spreading of ideas, knowledge and innovations.

The deterministic dynamical models are considered to be appropriate for the analysis of evolution of 'large' societal, scientific and technological systems for the case when the influence of fluctuations is unsignificant. Stochastic models are appropriate when the system of interest is 'small' but when the fluctuations become significant for its evolution. Finally the statistical approaches and models being based on the laws and distributions of Lotka, Bradford, Yule, Zipf - Mandelbrot, and others provide much useful information for the analysis of the evolution of systems which development is closely connected to the process of diffusion of ideas.

Knowledge and capital

- Knowledge can be defined as a dynamic framework connected to cognitive structures from which information can be sorted, processed and understood
- Capital is a collection of goods external to the economic agent that can be sold for money and from which an income can be derived.
- Often knowledge is parametrized as human capital

Knowledge is much more than a capital

- The concept of knowledge as a form of capital is an oversimplification. This (global-like) concept does not account for many properties of knowledge strictly connected to the individual, such as the possibility for different learning paths or different views, multiple levels of interpretations and preferences
- In fact, knowledge develops in a complex social context, and involves tacit dimensions requiring coding and decoding

Accumulation of knowledge

- The accumulation of the knowledge in a country population arises either from acquiring knowledge from abroad or from internal engines. The main engines for the production of new knowledge in a country are: the public research institutes, the universities and training institutes, the firms and the individuals. The users of the knowledge are firms, government, public institutions (such as the national education, health or security institutions), social organizations and the people. The knowledge is transferred from producers to the users by dissemination which is realized by flow or diffusion of knowledge .

Qualitative research. Science landscapes as examples for connection between qualitative and quantitative research

- The understanding of the diffusion of knowledge requires research complementary to mathematical investigations. For example, mathematics cannot indicate why the exposure to ideas leads to intellectual epidemics. Yet mathematics can provide answers to questions such as what is the intensity or the duration of some intellectual epidemics. Qualitative research is all about exploring issues, understanding phenomena, and answering questions. Qualitative research is empirical research in which the researcher explores relationships using textual, rather than quantitative data

Science landscapes

- The concept of science landscape is simple: Describe the corresponding field of science or technology through a function of parameters such as height, weight, size, technical data, etc. Then a virtual knowledge landscapes can be constructed from empirical data in order to to visualize and understand innovation and other processes in science and technology.
- One mathematical example of a technological landscape can be given by the function C=C(S,v) where C is the cost for developing a new airplane, S and v being the size and velocity of the airplane.

Application of science landscape for evaluating national research strategies

- The dynamics of self-organized structures in complex systems can be understood as the result of a search for optimal solutions on a certain problem. For example, national science systems can be considered as made of researchers which compete for scientific results, following optimal research strategies. The efforts of every country become visible, comparable and measurable by means of appropriate landscapes, e.g. the number of publications. The aggregate research strategies of a country can thereby be represented by the distribution of publications in the various scientific disciplines. In so doing, within a two-dimensional space, i.e., axes being by the scientific disciplines and the number of publications, different countries occupy different locations. Various political discussions can follow and evolution strategies invented thereafter

Lotka and Price – the pioneers

- Lotka – statistical approach. Lotka discovered a distribution for the number of authors n_r as a function of the number of published papers r

$$n_r = n_1 / r^2$$

- Price- deterministic model of scientific growth: Price distinguished three stages in the growth of knowledge:(a) a preliminary phase with small increments; (b) a phase of exponential growth; (c) a saturation stage. The stage (c) must be reached sooner or later after the new ideas and opportunities are exhausted; the growth slows down until a new trend emerges and gives rise to a new growth stage. The curve of this growth is a S-shaped logistic curve.

Population dynamics and epidemic models of knowledge diffusion

- Population dynamics is the branch of life sciences that studies short- and long-term changes in the size and age composition of populations, and how the biological and environmental processes influence those changes
- In the history of science and society, there are many examples of epidemic spreading-like of ideas. Examples of the former field pertains to the ideas of Newton on mechanics and the passion for "High Critical Temperature Superconductivity" at the end of the XX-th century. Examples of the latter field are the spreading of Moses or Buddha ideas , and discussions based on the Kermack-McKendrick model for the epidemic stages of revolutions or drug spreading.
- Epidemic models belong to a more general class of Lotka - Volterra models used in research on systems in the fields of biological population dynamics and social dynamics, but also economics, as well as for modeling processes connected to the spread of knowledge, ideas and innovations.

Large-size systems: deterministic models

- Logistic curve models:
- broadcasting model
- word-of-mouth model
- mixed model
- Influence of the time lag between the moment when a potential adopter hears about a new innovation and the time of adoption: Lotka-Volterra model

An example: Broadcasting model

Let us consider a population of K potential adopters of the new technology and let every of them switch to the new technology as soon as he/she hears about its existence (immediate infection). The probability that at time t a new subject will adopt the new technology is characterized by a coefficient of diffusion $\kappa(t)$ which might or might not be a function of the number of previous adopters. In the broadcasting model $\kappa(t) = a$ with $(0 < a < 1)$; this is considered to be a measure of the infection probability.

Let $N(t)$ be the number of adopters at time t. The increase in adopters for each period is equal to the probability of being infected multiplied by the current population of non-adopters [72]. The rate of diffusion at time t is

$$\frac{dN}{dt} = a[K - N(t)]. \tag{2}$$

The integration of (2) leads to the number of adopters, i.e.

$$N(t) = K[1 - \exp(-at)]. \tag{3}$$

Price model of knowledge growth. Cycles of growth of knowledge

- Price considered the exponential growth as a disease of science that retards the growth of stable science, producing narrower and less flexible specialists.
- Interesting result of the research of Price is that if a government wants to double the usefulness of science it has to multiply by about eight the gross number of workers and the total expenditure of manpower and national income

Price model of knowledge growth

- A generalized version of the Price model for the growth of a scientific field is based on the following assumptions:
- (a) the growth is measured by the number of important publications appearing at a given time
- (b) the growth has a continuous character though a finite time period T = const is needed to build up a result of the fundamental character
- (c) the interactions between various scientific fields are neglected. If in addition the number of scientists publishing results in this field is constant then the rate of scientific growth is proportional to the number of important publications at time t minus the time period T required to build up a fundamental result. The model equation is

$$\frac{dx}{dt} = \alpha x(t - T)$$

Models based on 3 or 4 populations

- SIR (Susceptible-Infected-Removed) model
- SEIR model for the spreading of scientific ideas
- SI discrete model of the change in the number of publications for a scientific field
- Discrete model for the population of papers (Daley model)
- Coupled discrete model for the populations of scientists and papers

An example: The SIR model

- Goffman and Newill considered as intellectual epidemics the stage of fast growth of scientific research in a scientific field and developed model of the epidemic stage of scientific research based on three classes of population: (i) the susceptibles S who can become infectives when in contact with infectious material (ii) the infectives I who host the infectious material (the ideas); (iii) the recovered R who are removed from the epidemia for different reasons.
- The epidemic stage is controlled by the system of equations

The SIR model (2)

$$\frac{dS}{dt} = -\beta SI - \delta S + \mu,$$
$$\frac{dI}{dt} = \beta SI - \gamma I + \nu,$$
$$\frac{dR}{dt} = \delta S + \gamma I$$

where μ and ν are the rates at which the new supply of susceptibles and infectives enter the population. A necessary condition for the process to enter the epidemic state is $\frac{dI}{dt} > 0$. Then

$$S > \frac{\gamma - \nu/I}{\beta} = \rho$$

The SIR model (3)

$$S > \frac{\gamma - \nu/I}{\beta} = \rho$$

is the threshold density of susceptibles, i.e. no epidemics can develop from time t_0 unless S_0 the number of susceptibles at that time, exceeds the threshold ρ. The epidemic state can not be maintained over a time interval unless the number of susceptibles is larger than ρ through that interval of time. As I increases ν/I converges to 0 and ρ converges rapidly to γ/β.

Continuous models of joint evolution of scientific sub-systems

- Coupled continuous model for the populations of scientists and papers: Goffman-Newill model
- Bruckner-Ebeling-Scharnhorst model for the growth of n subfields in a scientific field

Ideology of Bruckner-Ebeling-Scharnhorst model

- Three fundamental processes of evolution are included in the model equations:
- (a) self-reproduction: students and young scientists join the field of and start working on corresponding problems. Their choice is influenced mainly by the education process as well as by individual interests and by existing scientific schools
- (b) decline: scientists are active in science for a limited number of years. For different reasons (for example, retirement) they stop working and leave the system
- (c) field mobility: individuals turn to other fields of research for various reasons or maybe open up new ones themselves.

Small-size scientific and technological systems. Stochastic models

- Probabilistic SI and SEI models
- Master equation approach: (a) Stochastic evolution model with self-reproduction, decline and field mobility; (b) model of scientific productivity – in its simplest form can be reduced to a Fokker – Planck equation for the PDF for the age and productivity of the scientific community

Space-time models. Competition of ideas. Ideological struggle in a country with growing total population

$$\frac{\partial \rho_i}{\partial t} - D_{ij} \sum_{j=0}^{n} \Delta \rho_j = r_i \rho_i + \sum_{j=0}^{n} f_{ij} \rho_j + \sum_{j=0}^{n} a_{ij} \rho_i \rho_j + \sum_{j,k=0}^{n} b_{ijk} \rho_i \rho_j \rho_k + \dots,$$

$$\frac{\partial \rho}{\partial t} = r \rho \left(1 - \frac{\rho}{C}\right)$$

Mean field approximation – Verhulst-Lotka-Volterra model of ideological struggle

$$N = N_0 + \sum_{i=1}^{n} N_i; \qquad \frac{dN}{dt} = rN\left(1 - \frac{N}{C}\right)$$

$$\frac{dN_i}{dt} = r_i N_i + \sum_{j=0}^{n} f_{ij} N_j + \sum_{j=0}^{n} b_{ij} N_i N_j$$

Statistical approach

- Lotka – Volterra equations imply Pareto – Zipf distribution (Solomon and Richmond)
- Lotka's law. Distributions of Pareto and Yule.
- Pareto's distribution, Zipf - Mandelbrot' and Bradford's laws

Lotka's law (1926)

- Lotka noticed the following dependence for the number of scientists n_k who wrote k papers

$$n_k = \frac{n_1}{k^2}; \quad k = 1, 2, \ldots, k_{max}$$

- n_1 is the number of scientists who wrote just one paper and k_{max} is the maximal productivity of a scientist

Generalized law of Lotka

- Lotka's law is valid only on the average since the exponent in the denominator of equation above is not necessarily equal to two. Thus, Lotka's law should be considered as the most typical among a more general family of distributions

$$n_k = \frac{n_1}{k^{1+\alpha}}$$

- where alpha is the characteristic exponent of the distribution

Generalized law of Lotka

- The increase of alpha is accompanied by the increase of low productivity scientists.
- This means that when the total number of scientists is preserved and alpha increases the portion of highly productive scientists will decrease.
- Lotka's law is an asymptotic experssion for the Yule distribution of the scientific output

Yule distribution for the scientific output

- p(x): probability that an author has written x papers
- Yule distribution:

$$p(x) = \frac{\mu}{\lambda} B\left(x, \frac{\mu}{\lambda} + 1\right) = \alpha B(x, \alpha + 1), x = 1, 2, \ldots$$

where $B(x, \alpha+1) = \Gamma(x)\Gamma(\alpha x+1)/\Gamma(x+\alpha+1)$ is a beta-function, $\Gamma(x) \approx (x-1)!$ is a gamma-function, $\alpha = \mu/\lambda$ is the characteristic exponent. For instance if $\alpha \approx 1$ then $p(x) = 1/[x(x+1)]$. Let us assume that $x \to \infty$ and apply the Stirling formula. Thus the asymptotics of Yule distribution (70) is like Lotka' law (66) (to within a normalizing constant): $p(x) \propto \Gamma(\alpha+1)\alpha/x^{1+\alpha}$.

Pareto distribution: continuous analog of Lotka's law

For large enough values of the total number of scientists and the total number of publications we can make the transition from discrete to continuous representation of the corresponding variables and laws. The continuous analog of Lotka law (66) is Pareto distribution

$$p(x) = \frac{\alpha}{x_0}\left(\frac{x_0}{x}\right)^{\alpha+1}; \quad x \geq x_0; \quad \alpha > 0 \tag{71}$$

which describes the distribution density for a number of scientists with x papers; x_0 is the minimal productivity $x_0 << x << \infty$, a continuous quantity.

Zipf-Mandelbrot law: ranking the scientists with respect to their productivity

In order to obtain the law of Zipf-Mandelbrot we start from the following version of Lotka's : $n_x = C/(1+x)^{1+\alpha}$, where x is the scientist productivity, α is a characteristic exponent, C is a constant which in most cases is equal to the number of authors with the minimal productivity $x = 1$, i.e., to n_1. On the basis of this formula the number of scientists r who are characterized by productivity $x_r < x < k_{max}$ (k_{max} is the maximal productivity of a scientist) reads

$$r = \sum_{x=x_r}^{k_{max}} n_r \approx C \int_{x_r}^{k_{max}} \frac{dx}{x^{1+\alpha}} = \frac{C}{\alpha}\left(\frac{1}{x_r^{\alpha}} - \frac{1}{k_{max}^{\alpha}}\right). \qquad (72)$$

Depending on the value of x_r r can have values $1, 2, 3, \ldots$ and in such a way the scientists can be ranked. If all scientists of a scientific community working on the same topic are ranked in the order of the decrease of their productivity, the place of a scientist who has written x_r papers will be determined by his/her rank r. When the productivity of a scientist x_r is found from Eq.(72) as a function of rank r, the relationship

$$x_r = \left(\frac{A}{r+B}\right)^{\gamma}; \quad A = (C/\alpha)^{1/\alpha}; \quad B = C/(\alpha k_{\max}^{\alpha}); \quad \gamma = 1/\alpha. \qquad (73)$$

This is the rank law of Zipf-Mandelbrot, which generalizes Zipf law: $f(r) = cr^{-\beta}$; $r = 1, 2, 3, \ldots$, where c and β are parameters. Zipf law was discovered by counting words in books. If words in a book are ranked in decreasing order according to their number of occurrences then Zipf law states that the number of occurrences of a word is inversely proportional to its rank r.

Bradford's law: ranking the journals

Assuming that in Lotka law the exponent takes the value $\alpha = 1$ and that in most cases $C = n_1$, one has $x_r = n_1/(r + a)$ where $a = n_1/k_{max}$, $r \geq 0$. Integration of the last relationship yields the total productivity $R(n)$ of all scientists, beginning with the one with the greatest productivity k_{max} and ending with the scientist whose productivity corresponds to the rank n (the scientists are ranked in the order of diminishing productivity; the rank is assumed to be a continuous-like variable):

$$R(n) = n_1 \ln\left(\frac{n}{a} + 1\right) \tag{74}$$

This is Bradford law. According to this law, for a given topic, a large number of relevant articles will be concentrated in a small number of journals. The remaining articles will be dispersed over a large number of journals. Thus if scientific journals are arranged in order of decreasing published of articles

Bradford's law (2). Hierarchical stratification

on a given subject, they may be split to a core of journals more particularly devoted to the subject and a shell consisting of sub-shells of journals containing the same numbers of articles as the core. Then the number of journals from the core zone and succeeding sub-shells will follow the relationship $1 : n : n^2 : \ldots$.

The Zipf-Pareto law tells us that in the case of the distribution of scientists with respect to their productivity, one can always single out a small number of productive scientists who wrote the greatest number of papers on a given subject, and a large number of scientists with low productivity. The same applies also to scientific contacts, citation networks, etc. This specific feature (called hierarchical stratification) of the Zipf-Pareto law reflects a mechanism of formation of stable complex systems. This must be taken into account in the process of planning and organization of science.