Andrzej Jarynowski

Introduction to mathematical modelling

There is a huge variety of possible approaches to modelling. If we are interested in correspondence to reality, models can be precise, but very complicated. If we agree to lower the precision, models can be simple and easy to analyse. A model can be deterministic or stochastic; interactions can be implied by forces, energy, rules; variables can be discrete or continuous. Models may be further divided into two major, substantially different types, macroscopic and microscopic. In the case of macroscopic models, we want to answer the questions "how and how much". We do not care, what happens at the micro level of individual units of the analysis, only how the respective average values behave. Here, we are mainly dealing with all kinds of structural equations. This description is similar to the macroscopic description of complex systems, such as is the case in thermodynamics, which includes temperature, pressure, volume, etc. With this approach, you can answer a lot of quantitative questions; you can also generate more or less accurate predictions.

Macroscopic models and Malthus law case. The example is the Malthus model of population growth. Until mid-20th century, population growth on Earth was observed to be exponential. In the Malthus model, the world population keeps increasing (N) exponentially (curve "J"), while in the Verhulst model, it slows down, which is called a logistic curve ("S"). This also points to the very important issue of the reliability of models. The economists gathered around the Club of Rome predicted that human population would have exceeded 10 billion by 2015, while it is totally not the case. Not going into details, in which there are problems and traps in modelling.

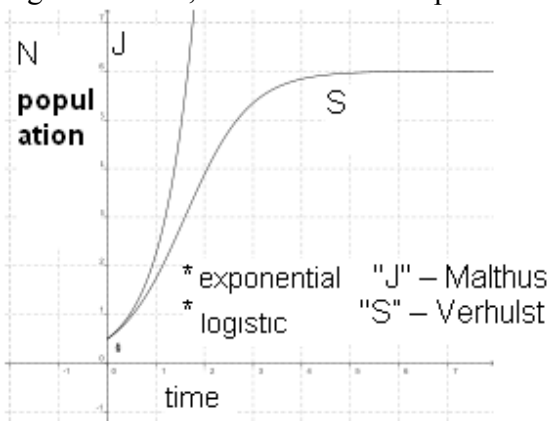

Fig. 1 The firsts models of population dynamics

Microscopic models. The main disadvantage of such models (macroscopic) is the lack of answers to the question about the causes for the occurrence of the phenomena ("why?"). Microscopic models were created in order to try to answer the question "why?". In the case of social and economic sciences, models are divided are divided as follows:

- Microsimulation, where the objects change their state due to deterministic or stochastic rules;
- Agent Based Models, wherein the system is a collection of "agents" interacting according to some dependent model rules.

An agent, as the basic element of the system, has some characteristics (described numerically) and usually interacts with other agents or external factors. The features of a single agent, as well as the rules, are affected depending on the specific model. You could say that building a named agent is a generalisation of the concept of particles, many body systems, etc. known from physics.

Modeling goals. The application of adequate theoretical methods and empirical approaches of rigorous sciences like mathematics and physics to economic and social issues has many faces. From the historical perspective, any quantitative science starts by collecting and systematizing empirical observations, then the search for regularities and patterns takes place, which finally results in theoretical formulation, which captures the observed behaviors and mechanisms. Although the current scientific community tries to make progress on all three stages, there are still methodological and conceptual issues that we think should be addressed in that context.

Software used in modeling and data analysis

Computational approach originated from the applications of engineering control systems needs software to be used. Inside this thesis:
- spreadsheets
- few langueages as R, C++ and Python;
- modeling enviroments as Vensim, Netlogo;
- Network Analysis Toolboxes as and Igraph.

VENSIM

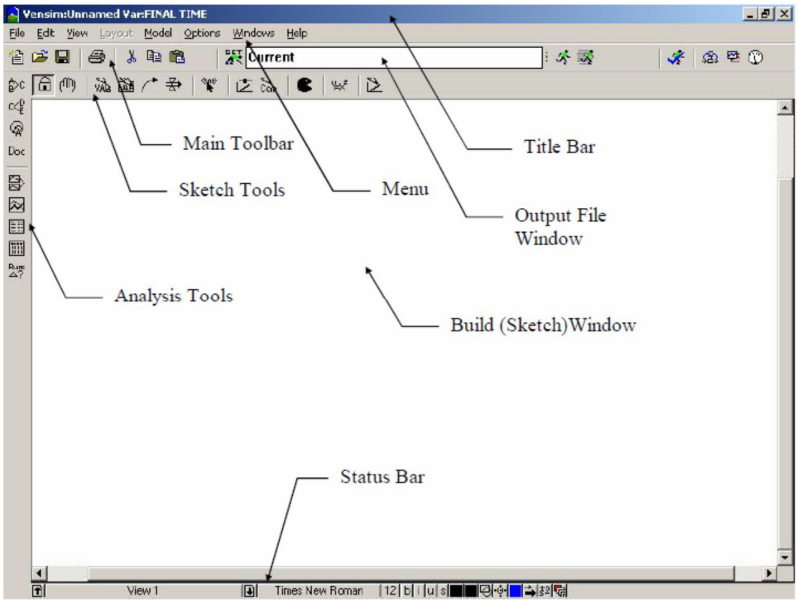

Fig. 2 User Interface of Vensin. Source: (Vensim 5 Modeling Guide, 2003)

In Vensim you can create your models by drawing items on the Sketch Window. The most important item (blocks) are:

- Variable- Auxiliary/constant (creates variables or constants). You can click on it to create a constant or a variable (auxiliary in Vensim terminology)
- Stock variable (creates stocks or levels). You can click on it to create stocks and work in the same way as for variables.
- Arrow tool creates arrows which will correspond to dependences
- Flow variable creates flows. For an inflow should click on the working area and drag to the stock variable. If the source for the flow is unknown, you start to draw the inflow by clicking in the working area and dragging to the stock. The same procedure is for the outflow with the only difference that you start at the stock and end in the working area.

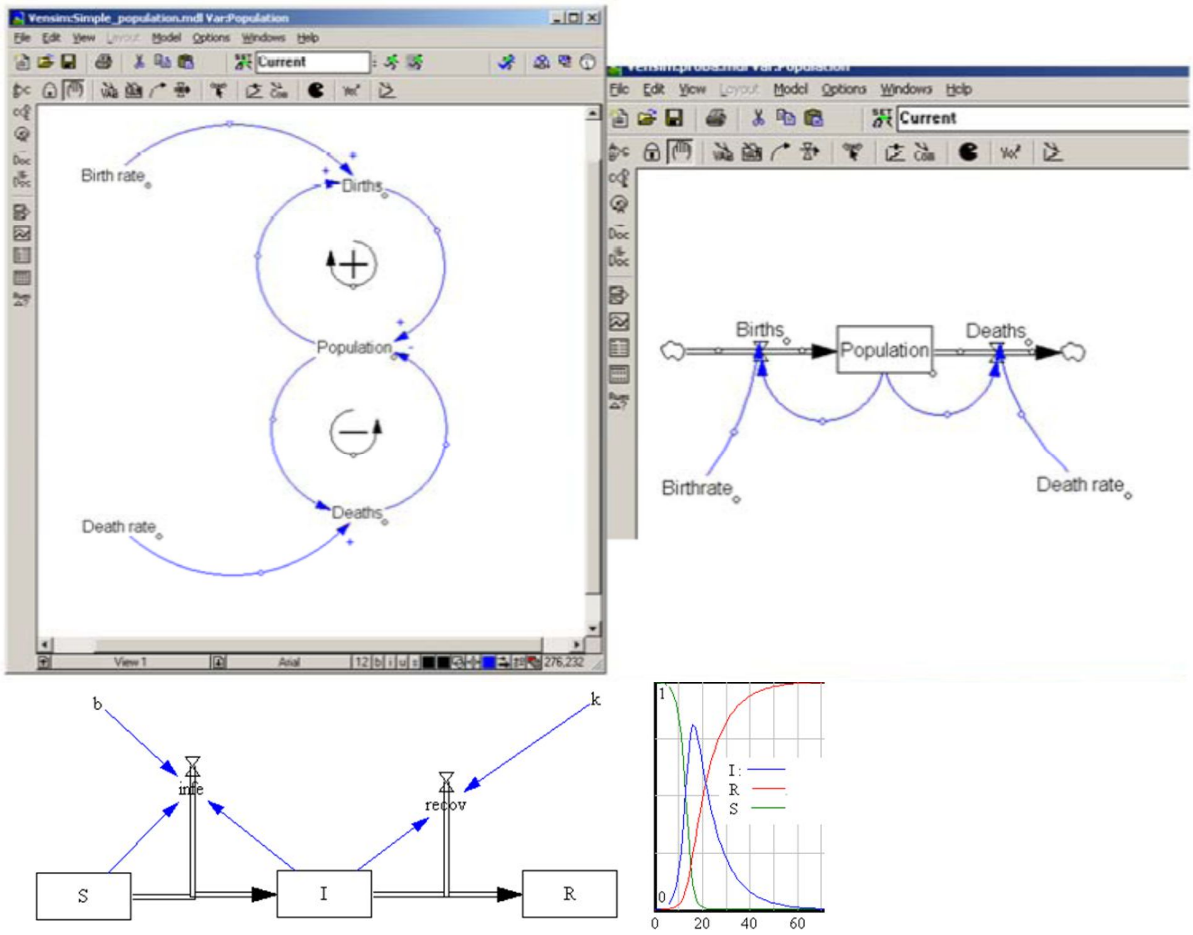

Fig. 2 Birth/Death process presented as Causal Diagram [top left] and Flow model [top right]. [bottom left] Susceptible, Infective, and Removed epidemic model (explained in details in the chapter on epidemiological modelling) in Vensim with parameters: *b*-infectivity, *k*-recovery rates and typical epidemic curve [bottom right]

NetLogo

The aim is to integrate a wide range of networks in which physical space is a crucial factor. I have used NetLogo for desctibing both spatial agents locations and movments with rule based characteristics of the model. Software cover the integration of empirical data and the automatic execution of multiple simulation runs, as well as the integration of ABM with spatial data (GIS).

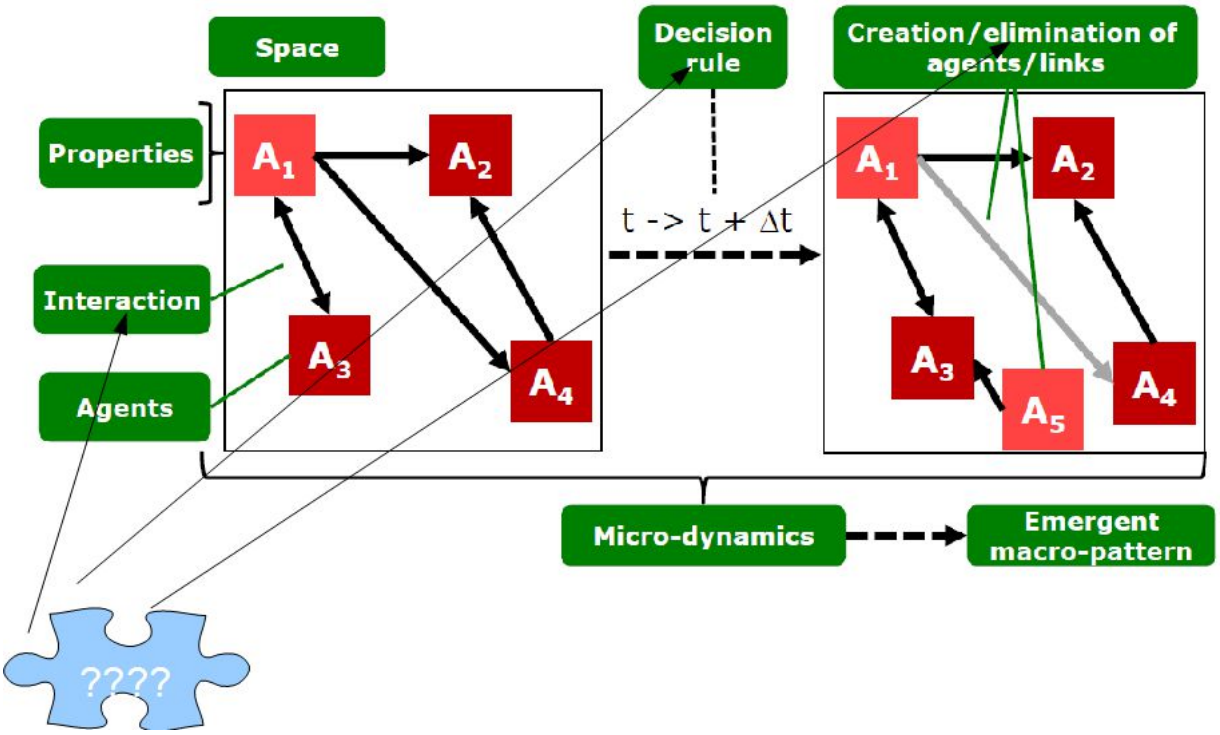

Fig. 3 General Agent-based methodology and the most difficult checkpoints wihin social phenomenon

In NetLogo GUI there is the three main tabs:

- Interface: where the model is run;
- Information: where the model is documented;
- Procedures: where the model is developed.

There are Four main types of user interface elements:

- The main model space (grey). It has origin and min/max x/y coordinates;
- Buttons to start/stop the model and to execute other procedures;
- Sliders/switches/choosers/inputs to enter parameters;
- Plots/output fields to retrieve information from the model.

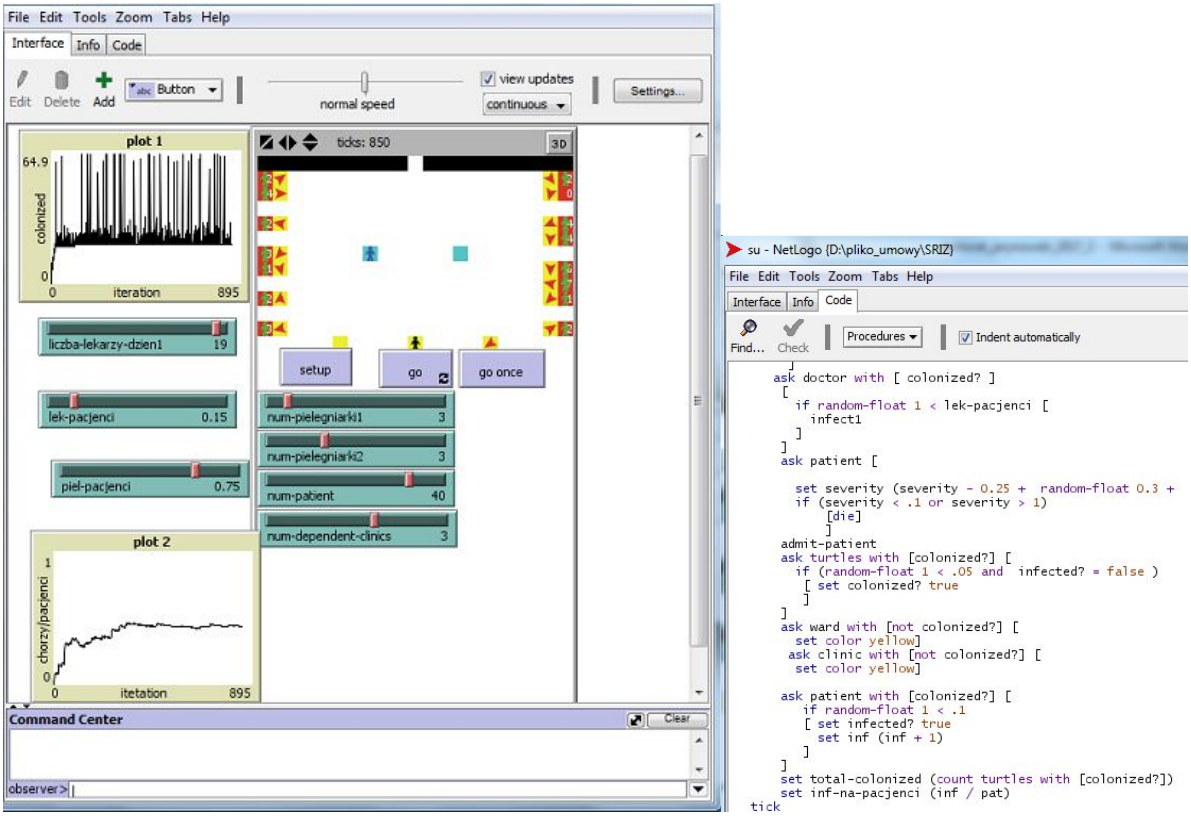

Fig. 4 GUI of Netlogo running a program and fragment of code from SIRS-Z project

NetLogo has a set of model agents with certain properties, which can run custom-written or built-in procedures ("primitives"). These agents are:

- Observer. An observer is an outside user of the model (user). The observer can run certain procedures directly, or "ask" other entities to run them.
- Turtles. A turtle is NetLogo's representation of an agent. It has certain built-in properties, but more can be added.
- Patches. A patch is the elementary spatial unit in the NetLogo grid. It has x/y coordinates, but also color etc. Other properties can be added.
- Links. A link is a connection between two agents.

In general, Netlogo is currently a tool widely used to simulate complex systems in social sciences. On the other hand, a there are plenty of languages or environments, to run stochastic processes whose also allow to manage spatial patters.

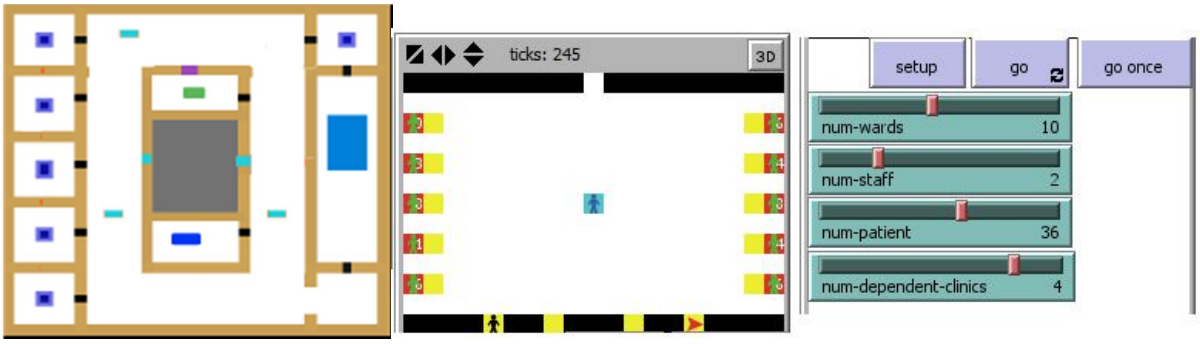

Fig. 5 Implementation of real buildings description and organizational habits of personnel in hospital infection project (SIRS-Z)

Epidemiological modeling

The area of epidemiological modelling is explored by researchers of different academic backgrounds: physicians, physicists, mathematicians, statisticians, computer scientists and sociologists. The mathematical modelling of the epidemiology of infectious diseases is an interdisciplinary science that supports public health institutions. All of them have something to add to this issue with different approaches, beginning from mathematics, through computer science (simulations) and concepts from physics, and ending with sophisticated sociological and statistical analyses. The aim of my work was initially to collect a few perspectives (mostly mathematical) and to develop them, but in effect the final analysis is more empirical, which makes it more practical in use. This work is also very extensive, because of its interdisciplinary nature, and highlights some methods that were not used in epidemiology until this moment. The epidemiological models that treat transmission as "human-to-human" from the differential equation point of view do exist in earlier literature, but in more recent agent-based models they appear more often. Mathematical models and computer simulation start to play a significant role as the quantity of social interactions is enormous, but real data, especially register-based, is more important than simulations. The first mathematical model in history describes the epidemiology of smallpox in Wroclaw. The author of that model, a famous mathematician named Daniel Bernoulli, presented his results in 1766 in Paris and concluded as follows, which is motto of my thesis: "I simply wish that, in a matter which so closely concerns the well-being of mankind, no decisions be made without all the knowledge that a little analysis and calculation can provide".

SIR model formulation within differential equations

The mathematical description of phenomena needs more assumptions to be made in terms of understanding the real situation. In the first instance, let us consider the spreading of a non-fatal disease, to which no person is naturally immune. Let us suppose that the population can be divided into two groups: the Susceptible and the Infectious.
Assume that at general time t:
$S(t)$ = Number of Susceptible
$I(t)$ = Number of Infectious
with $S(t)+I(t) = N$
The problem is to model spread of the disease.
Consider a single *susceptible* individual in a homogeneously mixing population. This individual contacts other members of the population at the rate C (with units' time^{-1}) and a proportion I/N of these contacts are with individuals who are infectious. If the probability of transmission of infection given contact is β, then the rate at which the infection is transmitted to *susceptible* is $\beta CI/N$, and the rate at which the *susceptible* population becomes *infected* is $\beta CSI/N$.

The *contact rate* is often a function of population density, reflecting the fact that the contacts take time and saturation occurs. One can envisage situations, in which C could be approximately proportional to N (which corresponds to mass action), and other situations, in which C may be

approximately constant. Hence, terms like βSI and $\beta SI/N$ are frequently seen in the literature. For these, and for many instances, in which the population density is constant, the contact rate function C is subsumed into β, which is now no longer a probability, but a "transmission coefficient" with units' time^{-1}. To reduce the value of the coefficient, let us write: $r=\beta C/N$. Thus, the development of the model allows for a possibility of recovery (an individual becomes immune to the disease). Looking at the recovery term, we assume that it is proportionally related to the infectious. In the end, let us suppose that population is divided into three classes: the susceptible (S), who can catch the disease; the infectious (I), who can transmit disease and have it; and the removed (R), who had the disease and have recovered (with immunity) or are isolated from society. The transition pattern can be represented as follows:

$$\frac{dI}{dt} = rSI - aI,$$
$$\frac{dS}{dt} = -rSI,$$
$$\frac{dR}{dt} = aI$$

A key question for the given r, a, S_0, I_0, is whether the infection will spread or not and if so, how it will develop in time and when it will start to decline. Since the initial condition for S- $S_0<a/r$ then $dI/dt<0$ in which case $I_0>I(t)$ and I goes to 0 with t going to infinity. On the other hand if $S_0>a/r$ then $I(t)$ increases and an epidemic appear. We have something like the threshold phenomenon S_c which depends on initial numbers. Concluding let me write:

$$R_0 = \frac{rS_0}{a}$$

where R_0 is basic reproduction rate of the infection. This rate is crucial for dealing with an epidemic which can be under control with vaccination for example. Action is needed if $R_0>1$, because then an epidemic clearly breaks out.

Reduction of SI model. In first instance let consider the spread of a non-fatal disease, to which no-one is naturally immune. Suppose the population can be divided into two groups: Susceptible-Healthy and Infectious-Infected. Such a model referred to a phase plane. The concern is only with values of I in the interval [0,N], indicating that I'(t)>0 which means that number of infectives is increasing. By inspection of figure it is seen that the arrow approaches the equilibrium state I=N for all permissible values of I. This means that the number of invectives will tend to the equilibrium state N, no matter how many invectives are initially present. Thus I(t)=0 is referred to as an unstable equilibrium state.

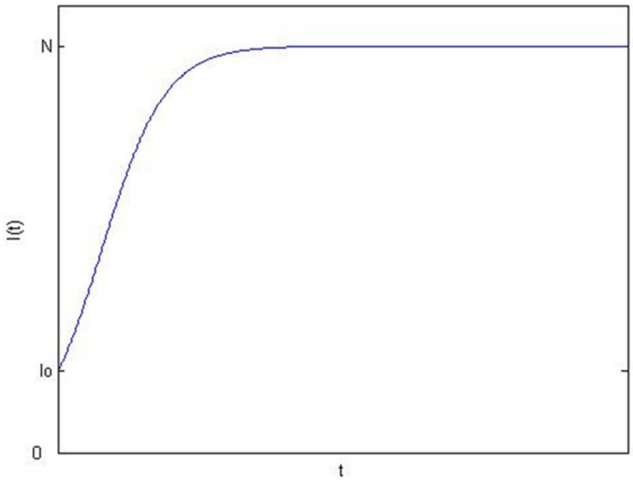

Fig. 6 Spread of Infection for SI model – where for a least on infected, whole population will be infected.

Extension of SIR model. The model can be expanded by an additional sub-population (e.g. Exposed to the pathogen - E -), or various kinds of pulses/delay functions to be more realistic with the actual processes. Although those non-linear extensions make the equation closer to the real problems, an analytical solution often cannot be developed (usually only an asymptotic or a solution with certain constraints is possible). Moreover, even the simplest SIR model presented above has no unique analytical solution for the whole parameter and initial condition space. Therefore, numerical methods are used - usually difference equations (a special case of the numerical procedure for obtaining the trajectory of the time).

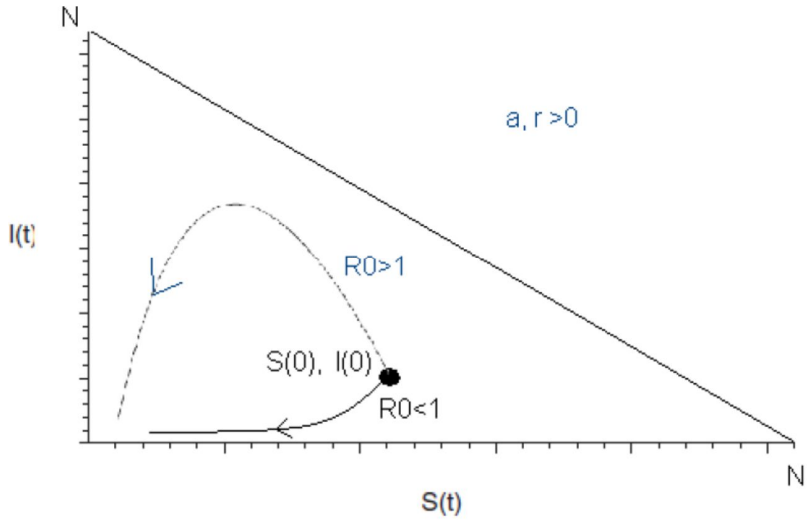

Fig. 7 The description of the stability of equilibrium points in simple SIR model, where no general analytical solution can be obtained

Other examples are partial differential equations in the geographic spread of diseases. The spatial component, e.g. the population density, can be applied in SIR models not only to describe the time evolution of the epidemic, but also of the transfer of the disease to new areas. For example, by slightly modifying the infecting equation to express the diffusion (where Δ is the Laplacian, and D is the diffusion coefficient), we obtain:

$$\frac{\partial I}{\partial t} = D\Delta I + rSI - aI$$

The dynamics of the spreading of the disease in a geographical space (diffusion) is used to map the epidemic of plague that started from the harbours of the Mediterranean Sea at the beginning of the 14th Century. It arrived to Poland five years later, went farther north and expired.

Stochastic model

A competitive approach to the deterministic modelling of an epidemic is the probabilistic method. Instead of model parameters, it uses rates, because we are dealing with probabilities. Therefore, for example, the rate of infection (movement from *S* to *I*) is represented by a pure probability of changing state, not as a flow between stocks, in the deterministic approach.
The main difference between the stochastic and the deterministic methodology is the meaning of R_0. In stochastic mode for $R_0<1$ it is already possible to start epidemic. Let us consider the Markovan discrete process with the notation ($I(t) = i$, $I(t + \Delta t) = j$):

$$p_{\Delta t}(i \to j) = \begin{pmatrix} ri(-i)\Delta t, & j = i+1 \\ ai\Delta t, & j = i-1 \\ 1-[ri(N-i)+ai]\Delta t, & j = i \\ 0, & other\ cases \end{pmatrix}$$

where $p(i->j)$ is the probability of state changing (size of cohort *I*).

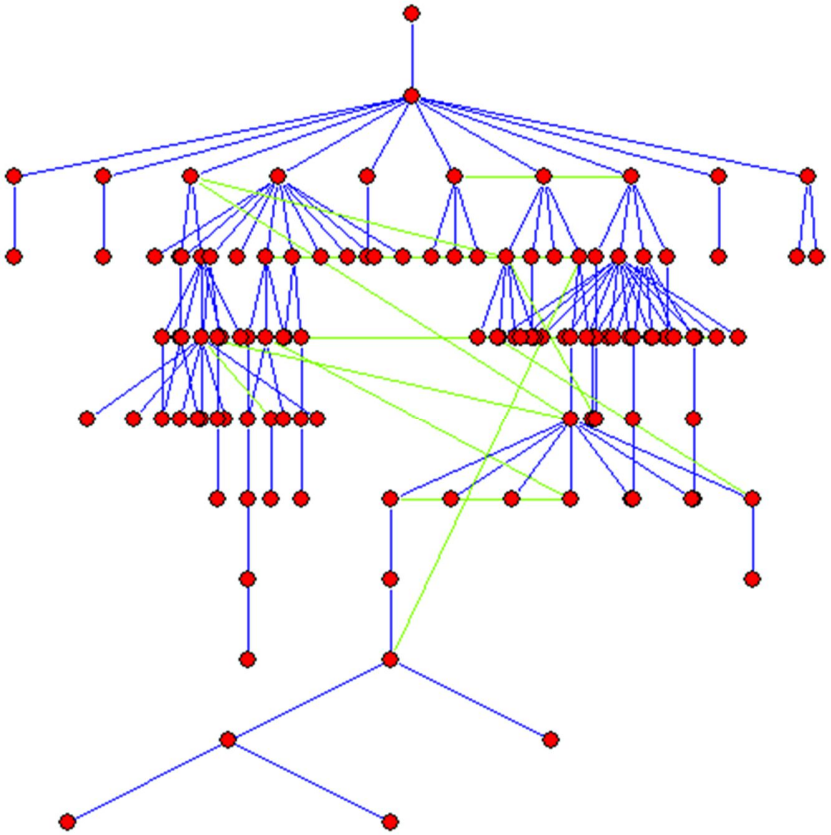

Fig. 8 The branching process of a real epidemic with 11 waves (generations). Blue is the primary and green the secondary (re) infections

In stochastic notation instead of rates, probabilities of changing state are more common. So β~r/N is probability of infection per time unit of of contact, and probability of transition between states I and R (recovery rate) equals γ~a. Epidemic reproduction rate can be written now in a new form: $R_0=\beta/\gamma$.

Cellular automata. In addition to the branching processes, cellular automata (grid cells with conditions and rules) are used in stochastic epidemiological modelling. The main objective of this approach is the statistical analysis of SIR models. Physicists also like those structures and many studies have been made of its topological characteristics, e.g. percolation. The epidemic is spreading through infected cells to neighbours in accordance with the established probabilistic rules. In addition to the standard models in the grid (most applications can be found in disease outbreaks in plants, where the system of flower beds literally has such a structure), there are all kinds of shortcuts corresponding to the vectors of infection.

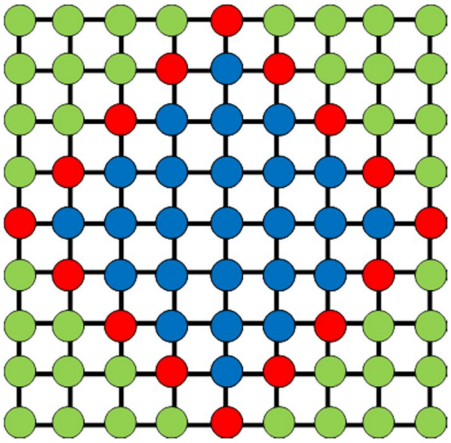

Fig. 9 Process of deterministic epidemic spread SIR started from single infective in central point of a grid. There are waves of infected - , after few steps of simulation change to removed- , and rest of population not yet affected susceptible - . It is CA with von Neumann neighborhood

Other propagation processes

The processes of information propagation as data flow are one of the core problems of complex systems and data science. Such pseudo-epidemic models can be of types like: diffusion of information (where mass is conserved), opinion formation, spread of influence or rumor spread.

State of the art in epidemiological modeling

Mathematical models can help in describing phenomena of epidemiology spreading and give an answer how to fight with them. Some tools are dedicated to epidemiologist and easy applicable, some are very sophisticated on theoretical level. Main goals of applied epidemiological modeling are to provide guidelines for controlling disease outbreaks. There are many problems in modeling because of big variety of epidemic types. Mathematical modelling of the epidemiology of infectious is an interdisciplinary science which supports
public health institutions. The use of the data and the use of computer simulation, in order to understand and modify the social processes in an innovative project are to personalized health care (knowledge of the location of patient on temporal network of contact could have an impact on decisions concerning the medical treatment). In my thesis, I focus on Sexually Transmittable Infections and Hospital Infections specific characteristics of such kind of spread will be described. Firstly, set of useful tricks will be presented here from some quick estimation methods through time series analysis up to real-life models.

R_0 post-epidemic estimation.
For SIR –like epidemics following approx. calculation for empirical reproduction rate calculation have been used within epidemiologist. The reproduction rate can be approximated for the SIR model from empirical data:

$$R_0 = \frac{-\ln\left(\frac{S_\infty}{S_0}\right)}{1-\frac{S_\infty}{S_0}}$$

Where S_0 is initial number of susceptible, S_∞ is number of people remaining susceptible after the outbreak.
The formula derivation is easy to obtain, but it is extremely powerful in its applications.

Early detection of outbreak - β estimation

The early growth estimation can be easily done by matching incidence trajectory to the exponential function, resulting in the approximation of the force of infection (β). With these calculations the basic reproduction rate of the epidemic (R_0), recovery/detection chance (γ) could be estimated. The early detection method in zeros approximation can be done by fitting the incidence curve to the exponential function, resulting estimation of the infectivity coefficient per time unit (β). According to simplified relation in SIR (Susceptible, Infectious, Removed) model $R_0=\beta/\gamma$, we check what range of possible parameters fitting the data are critical to satisfy epidemic condition ($R_0=1$).

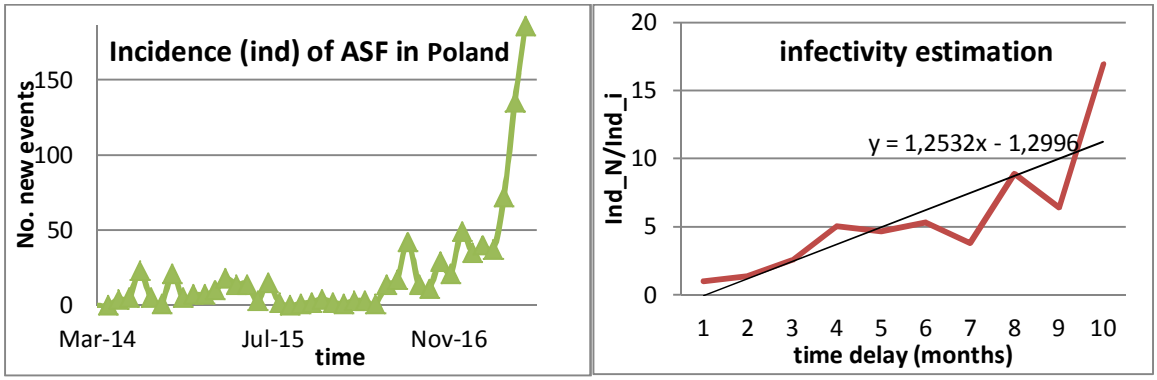

Fig. 10. ASF in Poland. Row incidence rates (Ind) [Left]. Fitting parameter β by incidence increments [Right].

Alert posting via time series analysis

Predictive modelling is the computational process by which a model is created or chosen to try to best predict the probability of an outbreak. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data, and for example new flu cases can be estimated based on previous season and current condition. There are many approach to modeling the incidence series with respect to seasonality and one of them has been implemented in R as a package – 'surveillance'.

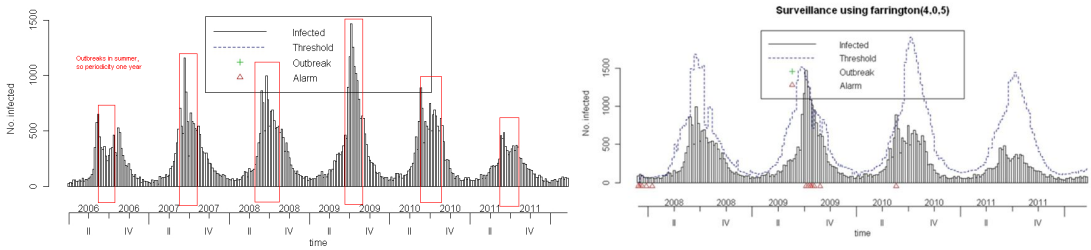

Fig. 11. Vomits queries per month in Sweden. Within 'surveillance' package view [Left]. Farrington algorithm for alarms [Right].

Presented package help epidemiologist to understand seasonal patterns and give some automatic alert if anomaly is detected.

Modelling coinfections. The spreading of co-infections (both mutualistic as well as antagonistic) in an empirical temporal network of contacts is an important challenge for modelers. Cooperative infections are very common in sexual infection and comorbidities are well risk factor for HIV. The same time cross-immunity phenomenon appears in HPV viruses. In bacterial hospital infection on the other hand, they often rule "there can be only one" appears and longnitudal studies do not show permanent co-infections in too many cases.

Cost-effectiveness and cost-benefit analysis

One of the most important topics of epidemiological modeling is the authoritative analysis of the costs and losses of potential epidemiological control strategies and identification of potential problems that health care will have to face in the future. Let consider more precise monetary index and measurement of population effects of intervention. To do so, epidemiologists introduce quality-adjusted life-year (QALY) - a measure of disease burden, including both the quality and the quantity of life lived The QALY is based on the number of years of life and adjusted to health state (where 1 is a perfect health and 0 is death) that would be saved by the intervention. Due to infection, QALY is decreased. The same time treatment as well as social costs of disease should be calculated. If there exist vaccination, which protect form infection, then compulsory, universal vaccination will be cost-beneficial, because such kind of intervention reduce the losses in QALY for dedicated population. However it could be too expensive respectively to national GDP if cost if intervention will be higher of cost of the disease. To classify, which intervention in cost-effective, the incremental cost per QALY yearly is below GDP per capita of given country or it could be partly cost-effective if it's below 3*GDP.

Complex networks and Social Network Analysis (SNA)

The network theory is useful when it comes to the study of nature from a systems perspective, and there are several examples, in which it has helped understanding the behaviour of complex systems. Genetic regulatory networks, Internet transfer protocols, social interactions and financial market dynamics are some examples, in which a network perspective is important to understand systems behavior. The most exciting property of these systems is the existence of emergent phenomena which cannot be simply derived or predicted solely from the knowledge of the

system' structure and the interactions between their individual elements. However, physics methodology proves helpful in many issues of complex systems properties including the collective effects and their coexistence with noise, long range interactions, the interplay between determinism and flexibility in evolution, scale invariance, criticality, multifractality and hierarchical structure. Thereby, complex networks are mostly artificial concept developed by physicists and mathematicians and (at least in theory) they obey universal rules. Complex network analysis not only helps better understand social behaviour and determine the degree to which individual agents build functioning and working system, but creates quantitative 'machine learning' approach for collective intelligence. However, a social network analysis technique has been used to support social theories in qualitative way by social scientists. The very concept of a social network was developed back in the 19th c. by Durkheim, who compared the structure and functioning of societies to biological systems consisting of interconnected components. In Poland, Bronisław Malinowski (1924-44) combined anthropological study with knowledge from the borders of psychology, mathematics and economics, trying to get a better grasp of the functioning of the world. Moreover Jacek Szmatka his team at Jagiellonian University in Cracow actively participated in 1990th in the development of Social Network Analysis as the one of first lab of this kind in Europe. Currently, SNA serves as a methodology and set of tools enabling a multifaceted in-depth exploration of interacting systems. A graph formally consists of set of nodes (vertices) V - v_i and set of links (edges) E - e_{ij}.

In network terminology:

- Node = an individual components of a network e.g. people, animals. In standard graph theory it has notation v_i.
- Edge = a direct or indirect link between components referred to in social networking e_{ij} as a relationship between two agents), could be weighted.
- Path = a route taken across components to connect two nodes. An example of a possible path is visualized on Fig. 9. Path searching has a long tradition in mathematical graph theory since 18th century Euler problem of bridges in Królewiec (Koeningburg).

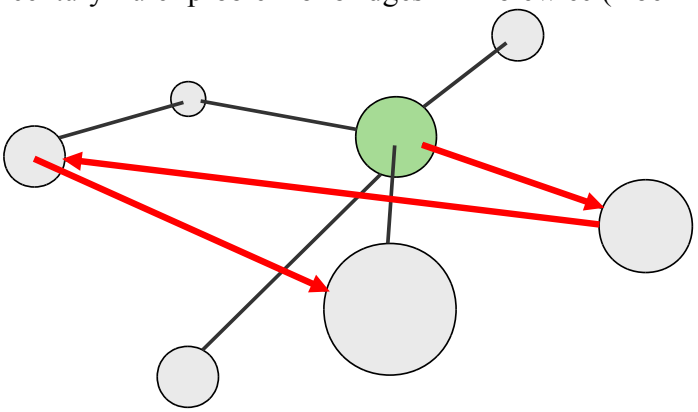

Fig. 12 Visualization of a path between nodes (arrowed links as a possible transmission channel

- Centrality = metric such as node degree, node centrality and betweenness centrality. By using some of these metrics it allow to measure both the number and strength of interactions and position of an individual.

To operate some practical properties, we must introduce some intermediate variables. The neighbourhood of a node *i* (N_i) is defined as its immediately connected neighbours. K_i is the

number of neighbours of a node (degree). Let consider, we have algorithms for find the shortest path.

Typical network properties are also listed:

- Community's structure. Community detection algorithms serve to automatically identify sub-groups within the observed population. A partition of network is a classification of the nodes that each node is assigned to exactly one of selected communities. Community is a social structure with connectivity within sub graph is higher than with the rest of the graph. There are several ways to make partitions and it is a computationally difficult task. Algorithms for finding communities can be of different type as: Minimum-cut method, Hierarchical clustering, Girvan–Newman algorithm, Modularity maximization, Statistical inference, Clique-based methods. In my studies, I choose just two very common community detection algorithms. Both, the Louvain Method and VOS Clustering for community detection are methods to extract communities based on function optimization once modularity once VOS quality.

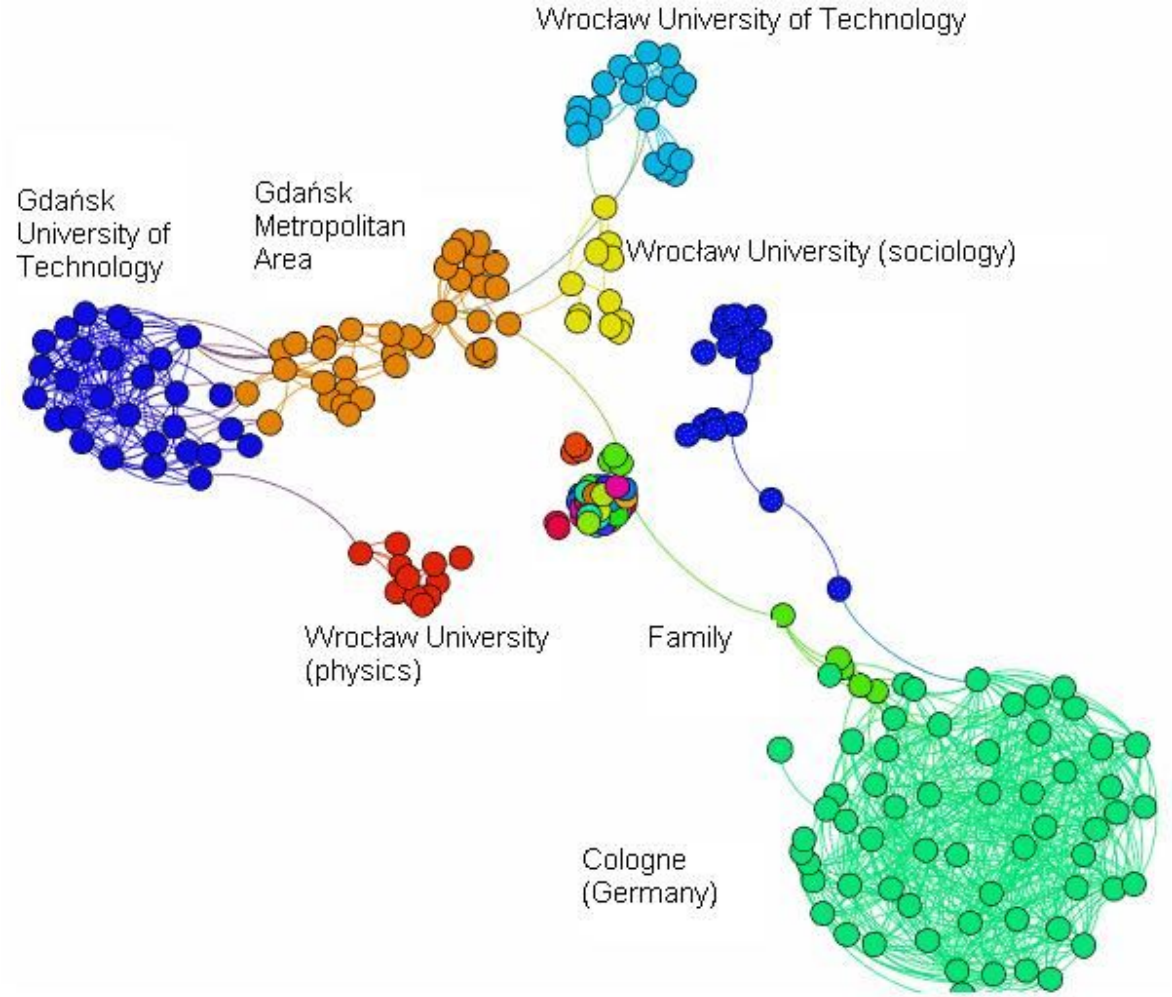

Fig. 16 Facebook network of author with annotated communities

- Average (shortest) path length – important for the flow of information in the network. Average shortest path length (L) in big enough real social was estimated as L=6 in S. Milgram experiment and 4.74 in Facebook analysis.
- Degree distribution- distribution of connections of nodes. There are some characteristic distribution (mainly long-tailed), which fit very well to empirical networks .

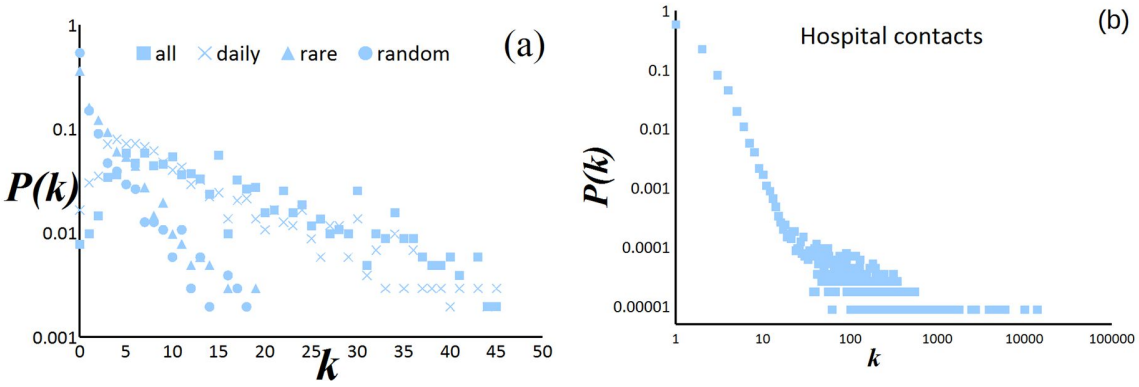

Fig. 14 Two characteristic degree distributions of in contact networks. (a) From POLYMOD contact survey for various contact loyalty used in study I.3.Polymod with exponential tail. (b) From Stockholm hospitals dataset with power-law tail.

- Assortativity: the assortativity coefficient is the correlation coefficient of degree between pairs of linked nodes.
- Clusters: A clustering coefficient counts the number of triangles in networks. Formally, a local clustering coefficient is defined by:

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

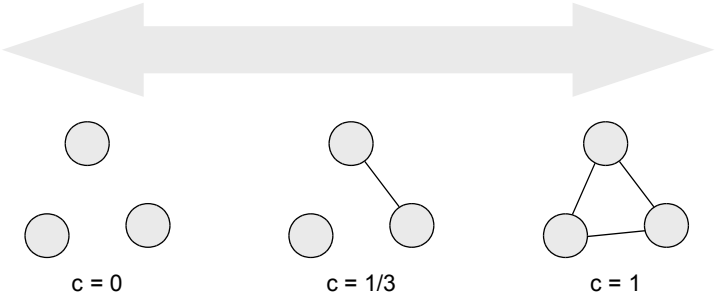

Fig. 15 Visualization for the clustering coefficient calculation

- Randomness: From a Grid/lattice network (ordered structure), via Small-world network (a mix of order and randomness), to Random networks (usually Barabasi-Albert: BA with power low degree distribution or Erdos-Renyi: ER with exponential degree distribution).

Classification and clustering. Partitioning as set of tools of exploratory data analysis that attempts to assess the interaction among patterns needs much more clarifying. If we do not know which nodes belongs to which category such kind of partitioning is called clustering – where within clusters nodes are closer to each other, than those whose patterns belong to different

clusters. “Groups” formed based on role of externally heterogeneous corresponds to classification procedure.

Dynamics and temporal aspects. However, those systems change in time at different scales and in different ways. The dynamics from this perspective have not been studied in detail and an integrative framework is missing

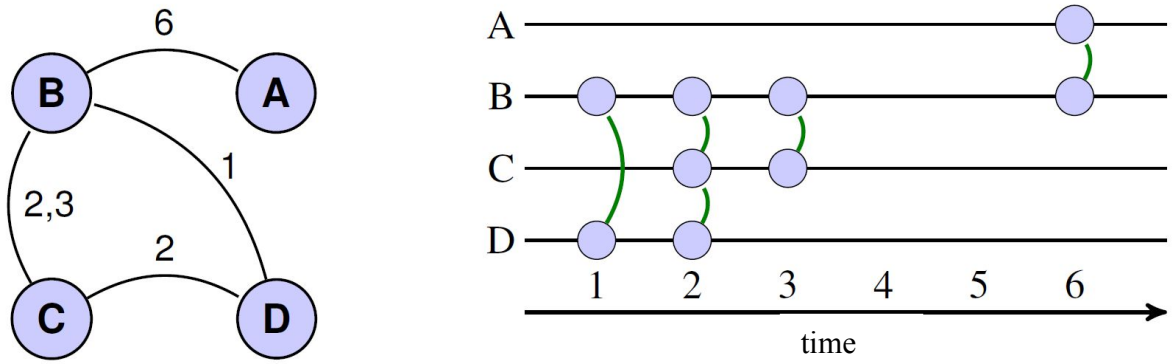

Fig. 16 Concept of temporal network, in which links are dynamic

Dynamical network models are important when we are dealing with complex systems, in which there are a significant number of nodes and interactions that show non-linear behaviour

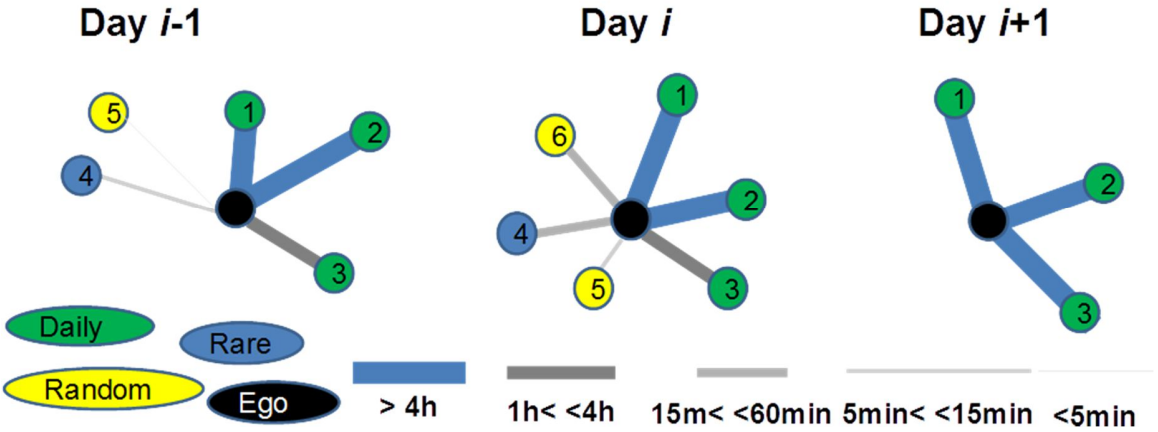

Fig. 17 The example of real evolving network. Time slices of ego network in POLYMOD study with different categories of intensity and loyalty of nodes