Model selection on networks

Matija Piškorec
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
matija.piskorec@irb.hr

***Abstract*—In this review paper we give a short overview of complex network theory and current state-of-the-art methods for statistical inference on networks. This includes methods for inference of network structure like communities, inference of links from data on network dynamics, and inference of processes on networks such as epidemics and social influence. Due to the variety of approaches and methods, which would be impossible to cover in detail, we are concentrating on those methods that are based on solid statistical foundations.**

Keywords: complex networks, dynamical systems on networks, Bayesian inference, model selection

I. Introduction

Complex network theory has its roots in a branch of mathematics called graph theory, which developed through the last two hundred years as a purely mathematical discipline. Historically, first applications of graph theory were in sociology for analysis of social networks and computer science for analysis of algorithms and data structures. In the last twenty years complex network theory experienced a great surge of interest from physics and computer science communities, borrowing many methods from statistical physics and theory of complex systems. This interest was facilitated by the increased availability of large amount of data on empirical networks such as Internet, World Wide Web, online social networks, metabolic and gene regulatory networks, and power grid networks. Potential applications in these domains helped transform complex network theory in a truly multidisciplinary field. Unfortunately, it is often the case that many phenomena on networks are explained by whatever model seems appropriate at a given time because of practical issues such as computation time or analytical tractability, instead of by validating proposed theoretical models against empirical data. We argue that this empirical validation is still crucial for the development of the field, and that more emphasis should be put on developing models which validate against empirical data. In the following sections we give an overview of complex networks theory, then present a framework for model selection based on Bayesian inference, and finish with current state-of-the-art methods for inference of models on networks.

Notes on terminology. There are several conventions on naming fundamental concepts that we are describing in this paper, depending on the community which uses them. *Graph theory* developed as a branch of mathematics in the middle of the 18^{th} century, and has been adopted in sociology, computer science and physics in the second half of the 20^{th} century. Due to the recent popularity of complex networks in general and online social networks in particular parts of these communities overlapped, leading to many inconsistencies in terminology. For example, mathematics community uses terminology *graphs*, *vertices* and *edges* while physics community uses *networks*, *nodes* and *connections* respectively. When it is important to highlight the ability of nodes to perform an action its usual to refer to them as *agents* (in physics) or *actors* (in sociology). When referring to the nodes in social networks it is usual to be specific and refer to them either as *persons* (in the case of real social networks) or *users* (in the case of online social networks). Similarly, *neighbors* are sometimes referred as *friends* in social networks. *Network panel data* is sometimes used in sociological literature as a synonym for *temporal networks*, and *digraph* is used for *directed network*. Naming of processes on networks depends largely on the context and the specificities of the process, with names like *diffusion*, *spreading*, *activation*, *rumor* and *cascade* being most common. In this paper we will mostly use terminology from physics while trying to be as specific as possible where necessary.

II. Fundamentals of complex network theory

Most common way [1] to represent an unweighted, undirected network is with a *adjacency matrix* A where element A_{ij} is 1 if there is a connection between nodes i and j, and 0 otherwise. It is also common to assume that there are no self-connection, that is $A_{ii} = 0$. We denote the total number of nodes in network with N. The number of connections that a node i has is its *degree* k_i, and the set of all nodes to which it is connected is called its *neighborhood*. The *degree distribution* P_k is a probability that a randomly chosen node from a network has a degree k. Classical Erdos-Reny random networks [2], where a connection between any pair of nodes has a probability p to exist, in the limit of a large number of nodes have a Poisson degree distribution $P_k = p^k e^{-p}/k!$. However, although it has many interesting properties, this model does not reproduce degree distributions of real networks, which are known to be *heavy-tailed* [3]. One particular type of a heavy-tailed distribution is a *power-law* [4] where $P_k \sim k^{-\gamma}$ as $k \to \infty$. Many generative models were proposed for networks with power-law degree distribution, including Price's model [5] and Barabasi-Albert model [6], that both depend in crucial way on *preferential attachment* of new nodes to the nodes with high degree. We will review

several methods of statistical inference of network evolution models in section V.

Many real world networks have a small *diameter* - the maximum length of the path that connects any two nodes in the network. There is evidence that diameter tends to stabilize or even shrink as the network grows larger over time [7]. Also, the proportion of connections in comparison to nodes tend to increase over time following a power law, a phenomenon known as *densification power law* [7]. Other distinct features of real networks are *clustering* on a microscale, *community* structure on a mesoscale, and *core-periphery* structure on a global scale. Global clustering coefficient measures a fraction of triangles - triplets of nodes that are connected to each other. It is much higher in real networks, especially in social networks, than Price's model and Barabasi-Albert model predict [1]. Local clustering coefficient [8] is defined for each particular node and it measures to what extent its neighborhood is close to a *clique* - network where each node is connected to each other. Communities in network [9] are groups of nodes which are more densely connected to each other than to other nodes in the network. The concept of a community is ill defined, resulting in variety of different definitions which result in different methods of their detection [10]. Choosing one particular definition usually means implicitly assuming a mechanism of community formation [11], so detecting communities requires a form of model selection. We will review methods for model selection regarding communities in section V.

III. Dynamical processes on networks

In this section we will give a brief overview of the current state of research on dynamical processes on networks, concentrating on the *contact processes* which are used to model biological and social contagions. We should note the difference between more general research area which deals with dynamical systems on networks, which covers dynamics of network structure as well as processes on networks. It would be impossible to comprehensibly cover the whole field, so we direct interested reader to the relevant review literature of dynamical systems on networks [12], as well as textbooks [13] and popular review articles [14], [15]. However, we believe that some methods used for inference of network structure, especially those based on solid probabilistic foundations or those that use data on dynamics (for example, from an information cascade in online social networks), could be useful for development of inference methods for dynamical processes on networks, and so we review them in section V-A.

A. Biological contagions

The most common way to study biological contagions in networks is through *compartmental models* [16] which describe possible states (or "compartments") of the nodes and rules which govern transition between the states. The most common compartmental models of biological contagion are *susceptible-infected* (SI), *susceptible-infected-susceptible* (SIS) and *susceptible-infected-recovered* (SIR), each defining corresponding compartments and transition rules which can be dependent on the number of infected neighbors (for example, when transitioning from susceptible to infected) or spontaneous (for example, when transitioning from infected to recovered, which is also an *absorbing* state). These three models are simple enough to be analytically tractable, with many theoretical results, but in order to gain insight into real epidemics one has to use more complicated compartmental models which are designed to be as realistic as possible and where parameters are estimated from real data [17], [18].

B. Social contagions

Methods for studying social contagions bear many similarities to the methods for studying biological contagions, most notably in the common usage of the phrase "contagion" for something which is essentially a social *influence*. There are decades of research originating from social science on social contagions [19], [20] and inference of dynamics in social networks [21], [22]. Bellow we briefly list some of the models of social influence, approximately in chronological order [12]:

- **Granovetter's threshold model.** [19] Also known as linear threshold model. A node is activated if the sum of influences from its neighbors exceed its own influence threshold.
- **Watts threshold model.** [20] Where each node can be in two states: inactive and active, and where each node has a threshold drawn from a distribution. Node is activated if the fraction of its activated neighbors exceeds its threshold.
- **Generalized model of contagion.** [23] Introduces the memory of past exposures which influences contagion, and can be used for both biological and social contagion. This model was motivated by the need to more finely distinguish between two extreme cases: (i) where successive contacts result in independent probability of infection, for example like in compartmental models and (ii) where there is a fixed threshold of contacts after which probability of infection immediately changes.
- **Centola-Macy model.** [24] Similar to Watts model, but uses absolute number of activated neighbors instead of their fraction.
- **Compartmental models.** [25] Inspired by biological contagions, compartmental models found their use in social contagions. Most notably *ignorant-spreader-stifler* (ISS) model which is similar to SIR model with a difference that transmission to absorbing state (stifler) is not spontaneous but depends on the presence of spreaders or stiflers in the neighborhood of the node.
- **Multiparametric model.** [26] Where activation of a node depends on the weighted linear combination of three terms: (i) personal preference, (ii) an average of its neighbors states and (iii) average of all nodes in the network.
- **Multi-stage complex contagions.** [27] Where nodes can be in one of three states: (0) inactive, where they exert no influence, (1) active, where they excert some influence

and (2) hyperactive, where they exert both regular and some bonus influence.

- **Synergistic model.** [28] Where infectivity and/or susceptibility of a node is dependent on the number of active neighbors.
- **Voter model.** [29] Where each node can be in one of two states, and at each time step one node is chosen uniformly at random from the network and it adopts uniformly at random a state from one of its neighbors.

Although there is research demonstrating that social contagions could be modeled with biological contagions [30], recent experimental evidence [31], [24] shows that social contagions have functional dependencies that are more complex than simple monotone dependency on the number of neighbors, as is the case in many biological contagion models. For example, instead of a person's number of neighbors, the parameter that drives the contagion is the number of connected components in the person's immediate neighborhood [31].

C. Influence and correlation in social networks

In order to measure social contagion, it is important to distinguish between correlation effects that arise in the network from the true influence (causation) from one person to another. First steps in this direction was done by using randomization strategies on network [32], which should diminish true influence and leave correlation intact. One form of correlation is homophily, which is a tendency of similar nodes to form connections between each other, and which is often confounded with the social contagion [33], [34], [35]. Moreover, there are factors outside networks that have an influence on contagion like political unrest [36], [37], natural disasters [38] and external media [39]. Also, the topics of the information themselves [40] can also be a significant driver for the social contagion.

IV. Model selection

The central goal of model selection is to select a model or a class of models which generalizes best to unobserved data [41]. In general this is not easy as the very features of the model that improve its fitness on the observed data, which is all we have during model selection, can actually decrease its fitness on the unobserved data. In statistics and machine learning community this is called *bias/variance* trade off, where models that consistently perform poor on both observed and unobserved data are considered to have high *bias*, and models that perform excellent on observed and poor on unobserved data are considered to have high *variance*. The crucial factor that determines the model's generalization performance is its *complexity*, which is a measure of the model's degrees of freedom which allow it to fit many different datasets. All model selection measures have to either explicitly or implicitly account for the complexity of the model. In section IV-B we will explain how properties of *marginal likelihood* (also known as *model evidence*) implicitly account for model's complexity in Bayesian model selection. In section IV-C we will describe several model selection measures which account for this complexity explicitly.

A. Relation between model selection and other related concepts

First we will comment on the subtle difference between two related concepts - *parameter estimation* and *model selection* [42]. Result of parameter estimation is always *one* specific hypothesis, which we will also call *point-hypothesis*. In comparison, model selection gives us a *set* of hypothesis. For example, a set of hypothesis could be a set of all k^{th} degree polynomials. One point-hypothesis from this set could be a k^{th} order polynomial with specific parameters. In general, we could also define a *model class* as a set of models with similar functional form, for example a model class of all polynomials of a model class of all Markov chains of arbitrary order. Here are several situations where model selection could prove more suitable than parameter estimation [42]:

- **When we want to select a "general theory".** For example, when we want to select a model that performs well under a variety of circumstances for which particular parameters of a model may differ.
- **Gaining insight.** After which we can perform more detailed experiments. This way we can investigate broad class of models with low precision first and then make more detailed analysis of this restricted class in order to find an optimal model or do parameter estimation.
- **Determining relevant variables.** By investigating subset of models which evaluates best under given data, and which probably share common relevant variables.
- **Prediction by weighted averaging.** Where we first find a model class, and then combine predictions made from point-hypothesis belonging to this model class.
- **Easier creation of model hierarchy.** If our model space has a hierarchical structure we can easily infer hierarchical models by evaluating subsets of models and composing them into hierarchy. This is not straightforward when doing parameter estimation.

Second, we will comment between the relation between model selection and *regularization*. The two concepts are essentially the same, as regularization also serves as a way to restrict the class of possible models in order to reach an unique solution to ill-posed problem or to prevent overfitting. From the Bayesian view, many regularization techniques correspond to imposing certain priors on the model's parameters. The distinction is mostly historical, although we must emphasize that regularization is often used in much broader context that includes parameter estimation along with model selection.

B. Bayesian model selection

In this section we will describe the problem of model selection from a Bayesian perspective [43], [44], [41] which gives probabilistically correct method of evaluating models while relying only on two basic rules of probability: (i) *sum rule* $p(X) = \sum_Y p(X, Y)$ and (ii) *product rule* $p(X, Y) = p(Y|X)p(X)$, where X and Y are two random variables,

$p(X)$ is a probability of a random variable X, $p(X,Y)$ is joint probability of random variables X and Y, and $p(Y|X)$ is a conditional probability of Y given X. Let us consider a set of L different models $\{\mathcal{M}_i\} = \{\mathcal{M}_1, \ldots, \mathcal{M}_L\}$ where each model $\mathcal{M}_i$ is defined as a probability distribution over an observed data set $\mathcal{D}$. Then we can evaluate the posterior distribution through *Bayes's theorem*:

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{\mathcal{M}_k \in \{\mathcal{M}_i\}} p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)} \tag{1}$$

Because the denominator is equal for all potential models $\mathcal{M}_i$, in model selection we can disregard it and reformulate the problem as:

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i) \tag{2}$$

The prior $p(\mathcal{M}_i)$ expresses a prior probability of different models before any data is observed. This probability could be uniform, or it could be chosen based on some characteristics of the models, for example their complexity [44]. *Model evidence* or *marginal likelihood* $p(\mathcal{D}|\mathcal{M}_i)$ expresses the preference for different models given the data. It can be viewed as a likelihood function over the space of the models in which the parameters θ have been marginalized out [41]:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta \tag{3}$$

Marginalization of the likelihood function $p(\mathcal{D}|\theta, \mathcal{M}_i)$ over the parameter space has an implicit effect of restricting the complexity of the model, and in this way preventing the *overfitting* of the model to the data [41]. Simple models concentrate their *probability mass* (in the case of discrete models, otherwise we should refer to the *probability density*) to the smaller number of datasets than complex models, but give each of them larger probability. This ensures that complex models will be penalized if data can indeed be explained by a more simple model [45].

The ratio of model evidences $p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$ for two models $\mathcal{M}_i$ and $\mathcal{M}_j$ is called *Bayes factor* [46]. We can show that Bayes factor is a ratio of posterior and prior odds [44]:

$$\underbrace{\frac{p(\mathcal{M}_i|\mathcal{D})}{p(\mathcal{M}_j|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}}_{\text{prior odds}} \tag{4}$$

In another words, Bayes factor quantifies the degree to which the newly observed data changed the evidence towards one or another model.

In the end, we should highlight the role of model evidence in evaluating posterior distribution over parameters for a particular model $\mathcal{M}_i$:

$$p(\theta|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \tag{5}$$

In this case the model evidence appears in the denominator and it acts as a normalizing constant.

C. Model selection criteria

It is hard to perform model selection in fully Bayesian treatment in the way described above because it requires integration of marginal likelihoods over all of the parameter space, which is often infeasible. Sometimes the maximum of the likelihood function is used as a first approximation to the model evidence, which is justified only if likelihood function is sharply peaked over maximum likelihood parameters. This is not true in general because it does not account for the shape of the likelihood function which is highly dependent on the functional form and the number of parameters of the model we are investigating. There are many approximate model selection criteria which use the maximum value of likelihood function in addition to different complexity terms in order to approximate model evidence. The most commonly used ones are Bayesian Information Criterion (BIC) [47], Akaike Information Criterion (AIC) [48] and Rissanen's Stochastic Complexity (SC) [49]:

$$\begin{aligned} \text{AIC} &= -2\ln f(y|\hat{\theta}) + 2k \\ \text{BIC} &= -2\ln f(y|\hat{\theta}) + k\ln N \\ \text{SC} &= -\ln f(y|\hat{\theta}) + \frac{k}{2}\ln N \end{aligned} \tag{6}$$

Where $y = (y_1, \ldots, y_N)$ is a data sample of size N, $\ln f(y|\hat{\theta})$ is the logarithm likelihood of the maximum likelihood parameters $\hat{\theta}$. Note that all of these measures feature first term which evaluates goodness-of-fit and the second term which evaluates model complexity, which depends only on the number of parameters of the model k and the number of observations N. When two models have the same number of parameters their comparison reduces to the generalized likelihood ratio testing [42]. But two different models could have the same number of parameters and still differ in complexity due to their functional form! To demonstrate this look at the following two models [50], [42]:

$$\begin{aligned} y &= ax^b + \text{error (Stevens' model)} \\ y &= a\ln(x+b) + \text{error (Fechner's model)} \end{aligned} \tag{7}$$

Although both models have the same number of parameters, Steven's model is more complex (in terms of the number of distributions it can fit, not the number of parameters), and it will always give a better fit to the data, making it more likely that it will fit noise along with data. So AIC and BIC criteria will consider Steven's model more general, and give it a higher score even in the case when data is generated from Fencher's model! So we need an additional complexity term that will properly account for the complexity of Steven's model that is due to the functional form. A measure of model selection which captures this is Minimum Description Length (MDL) [51], [52], [42]:

$$\text{MDL} = \underbrace{-\ln f(y|\hat{\theta})}_{\text{fitness}} + \underbrace{\frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int d\theta\sqrt{\det I(\theta)}}_{\text{complexity of a model family } f} \tag{8}$$

Where $I_{ij}(\theta)$ is the Fisher Information Matrix defined as the expectation value $I_{ij}(\theta) = -E_\theta[\partial^2 \ln f(y|\theta)/\partial\theta_i\partial\theta_j]$ evaluated in the distribution indexed with θ with a sample size of 1. MDL is the length in bits of the shortest possible code which describes the data generated by a model lying within the family f. MDL model selection is essentially the same as performing Bayes factor analysis with Jeffrey's prior [42].

D. Other model selection criteria

We will briefly review several other model selection criteria and their relation to the ones we described above.

Structural risk minimization (SRM) [53] uses similar trade-offs for model fitness (or "risk") and model complexity for model selection. It requires the definition of a nested space of models ordered by increasing complexity which is measured by *Vapnik-Chervonenkis dimension* (VC-dimension). In comparison to the above measures, the term for complexity is not in the same units as term for fitness, and so their combination is not straightforward [42]. SRM imposes no requirements on the type of models, as long as it is possible to calculate VC-dimension for them, and so the bounds it provides are very conservative, and can be considered as the *worst-case* estimate [41].

False Discovery Rate (FDR) [54] developed as a less conservative alternative for Bonferoni measure for multiple hypothesis testing. It works by controlling the expected proportion of rejected null-hypotheses which were in fact correct ("false discoveries"). In comparison to the measures above, which are based on Bayesian analysis, FDR is based on frequentist approach to statistics and it is useful only when there is a need to select one particular point-hypothesis out of a finite set of hypotheses.

Cross-validation (CV) [55] is a method where a model is repeatedly learned and tested on two disjoint subsets of the observed data, in hope to select for a model that will have good predictive accuracy on unobserved data. In this way model's complexity is incorporated implicitly because models that overfit on the training subset will be penalized by evaluation on the test subset. Under certain conditions, leave-one-out cross-validation (where there is only one sample in test dataset) asymptotically selects the same model as Akaike Information Criterion [56].

V. Statistical inference in networks

Over the years many methods for inference of network models were developed, many of which use some form of maximum likelihood estimation mentioned in section IV. Unfortunately, although evaluation of network models often employs statistical techniques in order to compare predictions of a model with empirical data, what is usually compared are only the *aggregated features* of the modeled and empirical networks like degree distribution or clustering coefficient in case of structure [57], or response correlations in case of dynamics [58]. Comparing only the aggregate features reduces the discriminative power of model validation [4], but is often practiced because it requires less computational resources and allows the usage of standard statistical methods. The advantage of using maximum likelihood for model selection is that different models can be compared directly in probabilistically unified way, rather than through the agreement of their predictions with a selected subset of many possible aggregated features [59].

There are several issues which have to be accounted for in order to perform statistical inference of network structure directly, rather than just the statistical comparison of their aggregated features [60], [61]:

- **Granularity of observations.** In comparison to standard statistical problems, a realization of a network generated by a model is considered as a single observation instead of a set of independent, identically distributed observations. This prevents us of using model selection methods which depend on partitioning the data set into independent training and test sets, such as cross-validation, because this would be impossible to do on networks. In reality this is not such a problem because a likelihood function anyway measures how a model predicts the *entire* data set, in our case the observed network. So instead of cross-validation we can use model selection methods which evaluate model fitness and model complexity without the use of independent test set, some of which we described in section IV-B.
- **Node correspondence.** This problem stems from the fact that a particular labeling of the nodes in network should not affect the likelihood function, as *isomorphic* networks should have the same likelihood of being generated by any particular model. So in order to calculate a likelihood we have to consider each of $N!$ possible permutations of node labellings, which is computationally infeasible. Estimation of likelihoods could be made much more efficient by using appropriate sampling strategies, for example Markov Chain Monte Carlo (MCMC) [60].
- **Likelihood estimation.** Even without the node correspondence problem,in order to calculate the likelihood that a particular model generated the observed network we still need to evaluate the probability of each of the N^2 possible edges in the observed network. Again, estimation of likelihoods could be made much more efficient by using appropriate sampling strategies like MCMC.

In the following two subsections we will review some of the methods used for inference of network *structure* (section V-A) and inference of *processes* on networks (section V-B). As we already mentioned in section III, we decided to distinguish between models which implicitly or explicitly use network dynamics for inference of network structure and models of processes on networks. Former include, for example, network growth models, models of community formation, and models of network structure inferred from dynamic data such as information cascades in online social networks. Later include, for example, epidemic and birth-death processes, biochemical and regulatory dynamics, human trails on the Web such as Web navigation and sequences of reviews.

A. Inference of network structure

Historically, sociological studies on human social networks predate most of the research on complex network structure and dynamics [62]. They are more concerned with modeling the dynamics of individual nodes, rather than modeling dynamics of a network on a global scale. For example, *actor models* are used to model the conditions under which nodes change their outgoing connections [21]. There are efficient maximum likelihood methods which are able to infer actor models from empirical data on network dynamics, and which incorporate many sociologically relevant features such as *transitive triplets*, *reciprocated ties*, *indirect ties* and *persistent reciprocity* [22].

Network growth models are concerned with the evolution and properties of network's global structure, and the problem of their inference gained a lot of attention in the research community [60], [59], [61], [63]. One of the first statistically principled approaches was to recognize that a growth model is actually a probability distribution on a space of all possible networks and to use maximum likelihood estimation to obtain most probable model giving the data on network growth [60]. This was also the first approach that used efficient Markov Chain Monte Carlo algorithm for the estimation of likelihoods. Maximum likelihood was also used to design complex models of network growth that are composed out of simpler microscopic principles [59]. We can reduce the computational cost of likelihood estimations by using less data, which is usually discriminative enough for model selection as compared to parameter estimation [63]. Also, supervised machine learning models which used network's aggregated features were used to discriminate between networks generated with different generative models, and to estimate their parameters [64].

Two network models which are often used for the representation of *network structure*, and for which efficient inference methods were developed are *Kronecker graphs* [61] and *block models* [65].

Kronecker graphs [61] are recursive models of networks that are expressive enough to model real network and to reproduce most of their properties. They rely on the *kronecker product* of adjacency matrices A and B which is defined as:

$$A \otimes B = \begin{pmatrix} a_{1,1}B & \cdots & a_{1,n}B \\ \vdots & \ddots & \vdots \\ a_{m,1}B & \cdots & a_{m,n}B \end{pmatrix} \tag{9}$$

Starting with an *initiator graph* K_1 and by successively applying kronecker product $K_1^k = K_1 \otimes \cdots \otimes$ we can generate self-similar network of arbitrary size. The Kronecker graphs have a multinomial distribution for in and out degrees of the nodes, which for some choices of initiator graphs behaves like a power-law distribution. Also, they follow the densification power law.

Another suitable representation of network structure are block models [65]. A block model that contains k blocks (or communities) is a $k \times k$ matrix M where each element M_{ij} gives a probability that a node from block k_i is connected to a node from block k_j. Erdös-Reny networks are a special case of block models where there is only one block. Minimum description length can be used in blockmodel inference as a complexity measure [66]. It is possible to infer block models from data in a *nonparametric* way, without predefining number of blocks [67]. Also, there are efficient Monte Carlo methods for inference of block models that optimize entropy rather than log-likelihood, and which can perform inference in a hierarchical way where every level serves as a prior information for the lower level [11].

Sometimes the information on network structure is lacking and the information on dynamics on network is used to infer it. For example, *CoNNIe* [68], *NetRate* [69], *NetInf* [70] and *InfoPath* [71] algorithms use generative probabilistic models for inferring network structure from information diffusion data. They all aim to find a spreading cascade which maximizes the likelihood of the observed data [72]. For this they use different optimization methods - convex programming for CoNNIe and NetRate, submodular function optimization for NetInf and stochastic gradients for InfoPath. Only InfoPath is able to provide an online estimate in case when network is changing over time.

B. Inference of processes on networks

In comparison to inference of network structure, inference of processes on networks still lacks a suitable representation which would allow inference of a broad range of dynamical models using an unified probabilistic framework [73]. In case of binary-state dynamics, where each node can occupy one of two states, we can use *infection rate* $F_{k,m}$ and *recovery rate* $R_{k,m}$ functions which depend only on the degree of node and the number of its neighbors, and which can describe many binary-state processes like SI and SIS models, Bass and Kirman models and voter models. These rate functions can be used to derive a *master equation* for describing time evolutions of the fractions of nodes in each of the states [74]. Unfortunately, currently there are no proposed methods for inference of these functions from data.

Another suitable representation is a general network dynamics equation [13] given in the form:

$$\frac{dx_i}{dt} = M_0(x_i(t)) + \sum_{j=1}^{N} A_{ij} M_1(x_i(t)) M_2(x_j(t)) \tag{10}$$

where x_i is an activity of node i, A is an adjacency matrix of the network, and nonlinear functions $M_0(x)$, $M_1(x)$, $M_2(x)$ define a space of dynamical models which has to be inferred. Examples of dynamical processes which can be represented with this equation are (i) epidemic processes, where x_i represents probability of infection of a node i, (ii) biochemical dynamics, where x_i represents concentration of a reactant i, (iii) birth-death processes, where x_i represents population at site i and (iv) regulatory dynamics, where x_i represents expression level of a gene i [73]. We can expand each function $M(x)$ into *Hahn series*:

$$M(x) = \sum_{n=0}^{\infty} A_n (x_0 - x)^{\Pi(n)} \tag{11}$$

which is a generalization of the Taylor's expansion that includes both negative and real powers $\Pi(n)$. The leading term of this expansion, which corresponds to power $\Pi(0)$, gives first approximation to the functional form of the dynamic process. It can be inferred using aggregated features of the *transient response* $x_i(t)$, which describes a response of the system after perturbation, and the *response matrix* G_{ij}, in this way finding a minimal model for the dynamical process on network [58]. This minimal model describes only the functional form of the model, and does not rely on model parameters A_n.

Human navigation on the Web can be modeled by Markov chain model, where Web content like Web sites, multimedia and reviews are states and sequences (or "trails") from one content to another are governed by transition probabilities [75]. There are efficient Bayesian inference methods which allow selection between prespecified hypotheses expressed as Markov chains from empirical data [76], [77].

There has been much research on the prediction of information cascades in networks [72] given past diffusion traces. First class of methods uses explicit information on network structure. Linear threshold model [19] can be fitted to data using gradient ascent method [78], although this can not reproduce realistic temporal dynamics [72]. *AsIC* and *AsLT* are asynchronous versions of independent cascades model and linear threshold models, and they can be inferred from data using a maximum likelihood estimation [79]. *T-BaSIC* model (Time-Based Asynchronous Independent Cascades) uses logistic regression to estimate functions depending on time which serve as model parameters [80]. Second class of methods do not assume existence of specific graph structure. For example, a SIS model can be fitted to data under assumptions that all nodes have the same probability to adopt the information and that they become susceptible at the next time step [81]. *Linear Influence Model* relaxes this assumptions, and it allows inference individual influence functions for each node separately in a non-parametric way by solving a non-negative least squares problem using the Reflective Newton Method [82]. *Partial Differential Equation* based model can predict topological and temporal dynamics of an information injected in the network by a given node, and its parameters can be estimated using the Cubic Spline Interpolation method [83].

ACKNOWLEDGMENT

This review paper has been written for the course: "Research seminar in the Computer Science" in the first year of postgraduate study 2014./2015. at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. October 2015. I would like to thank my mentors Mile Šikić and Tomislav Šmuc for valuable discussion.

REFERENCES

[1] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[2] P. Erdos and A. Reny, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[3] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[4] M. Stumpf and M. Porter, "Critical truths about power laws," *Science*, vol. 335, no. 6069, pp. 665–666, 2012.

[5] D. D. S. Price, "A general theory of bibliometric and other cumulative advantage processes," *Journal of the American Society for Information Science*, vol. 27, no. 5, pp. 292–306, 1976.

[6] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[7] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ser. KDD '05, 2005, pp. 177–187.

[8] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[9] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[11] T. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, 2014.

[12] M. A. Porter and J. P. Gleeson, "Dynamical systems on networks: A tutorial," *arXiv:1403.7663*, 2014.

[13] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical processes on complex networks*, 2008.

[14] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nature Physics*, vol. 8, no. 1, pp. 32–39, 2012.

[15] A. Motter and R. Albert, "Networks in motion," *Physics Today*, vol. 65, no. 4, pp. 43–48, 2012.

[16] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*. Springer Verlag, 2012.

[17] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions," *PLoS Med*, vol. 4, no. 1, 01 2007.

[18] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani, "Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility," *BMC medicine*, vol. 7, no. 45, 2009.

[19] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, vol. 83, no. 6, p. 1420, 1978.

[20] D. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 5766–5771, 2002.

[21] T. Snijders, G. van de Bunt, and C. Steglich, "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, vol. 32, no. 1, pp. 44–60, 2010.

[22] T. Snijders, J. Koskinen, and M. Schweinberger, "Maximum likelihood estimation for social network dynamics," *Annals of Applied Statistics*, vol. 4, no. 2, pp. 567–588, 2010.

[23] P. Dodds and D. Watts, "A generalized model of social and biological contagion," *Journal of Theoretical Biology*, vol. 232, no. 4, pp. 587–604, 2005.

[24] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.

[25] Y. Moreno, M. Nekovee, and A. Pacheco, "Dynamics of rumor spreading in complex networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 6 2, pp. 066 130–1–066 130–7, 2004.

[26] N. J. McCullen, A. M. Rucklidge, C. S. E. Bale, T. J. Foxon, and W. F. Gale, "Multiparameter models of innovation diffusion on complex networks," *SIAM Journal on Applied Dynamical Systems*, vol. 12, no. 1, pp. 515–532, 2013.

[27] S. Melnik, J. Ward, J. Gleeson, and M. Porter, "Multi-stage complex contagions," *Chaos*, vol. 23, no. 1, 2013.

[28] F. Pérez-Reche, J. Ludlam, S. Taraskin, and C. Gilligan, "Synergy in spreading processes: From exploitative to explorative foraging strategies," *Physical Review Letters*, vol. 106, no. 21, 2011.

[29] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of Modern Physics*, vol. 81, no. 2, pp. 591–646, 2009.

[30] A. Hill, D. Rand, M. Nowak, and N. Christakis, "Infectious disease modeling of social contagion in networks," *PLoS Computational Biology*, vol. 6, no. 11, 2010.

[31] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 16, pp. 5962–5966, 2012.

[32] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," 2008, pp. 7–15.

[33] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21 544–21 549, 2009.

[34] C. Shalizi and A. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociological Methods and Research*, vol. 40, no. 2, pp. 211–239, 2011.

[35] T. La Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," 2010, pp. 601–610.

[36] J. Borge-Holthoefer, A. Rivero, I. García, E. Cauhé, A. Ferrer, D. Ferrer, D. Francos, D. Iñiguez, M. Pérez, G. Ruiz, F. Sanz, F. Serrano, C. Viñas, A. Tarancón, and Y. Moreno, "Structural and dynamical patterns on online social networks: The spanish may 15th movement as a case study," *PLoS ONE*, vol. 6, no. 8, 2011.

[37] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "The dynamics of protest recruitment through an online network," *Scientific Reports*, vol. 1, 2011.

[38] X. Lu and C. Brelsford, "Network structure and community evolution on twitter: Human behavior change in response to the 2011 japanese earthquake and tsunami," *Scientific Reports*, vol. 4, 2014.

[39] W. Quattrociocchi, G. Caldarelli, and A. Scala, "Opinion dynamics on interacting networks: Media competition and social influence," *Scientific Reports*, vol. 4, 2014.

[40] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," 2009, pp. 807–815.

[41] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.

[42] P. Grunwald, *The Minimum Description Length Principle*. The MIT Press, 2007.

[43] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis, 3rd Edition*. Chapman and Hall/CRC, 2013.

[44] T. Ando, *Bayesian Model Selection and Statistical Modeling*. Chapman and Hall/CRC, 2010.

[45] I. Murray and Z. Ghahramani, "A note on the evidence and bayesian occam's razor," Gatsby Computational Neuroscience Unit, University College London, Tech. Rep., 2005. [Online]. Available: http://mlg.eng.cam.ac.uk/zoubin/papers/05occam/occam.pdf

[46] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, 1995.

[47] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[48] H. Akaike, "New look at the statistical model identification." *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716–723, 1974.

[49] J. Rissanen, "Universal coding, information, prediction, and estimation." *IEEE Transactions on Information Theory*, vol. IT-30, no. 4, pp. 629–636, 1984.

[50] I. Myung, "The importance of complexity in model selection," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 190–204, 2000.

[51] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[52] P. Grunwald, "Model selection based on minimum description length," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 133–152, 2000.

[53] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1989.

[54] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerfull approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[55] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.

[56] ——, "An asymptotic equivalence of choice of model by cross-validation and akaike's criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.

[57] F. Papadopoulos, M. Kitsak, M. Serrano, M. Boguñá, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[58] B. Barzel, Y.-Y. Liu, and A.-L. Barabási, "Constructing minimal models for complex system dynamics," *Nature Communications*, vol. 6, 2015.

[59] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," 2008, pp. 462–470.

[60] I. Bezáková, A. Kalai, and R. Santhanam, "Graph model selection using maximum likelihood," vol. 2006, 2006, pp. 105–112.

[61] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *Journal of Machine Learning Research*, vol. 11, pp. 985–1042, 2010.

[62] L. C.F., *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.

[63] M. Medo, "Statistical validation of high-dimensional models of growing networks," *Physical Review E*, vol. 89, 2014.

[64] S. Aliakbary, S. Motallebi, S. Rashidian, J. Habibi, and A. Movaghar, "Noise-tolerant model selection and parameter estimation for complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 427, pp. 100–112, 2015.

[65] P. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[66] M. Rosvall and C. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7327–7331, 2007.

[67] T. Peixoto, "Parsimonious module inference in large networks," *Physical Review Letters*, vol. 110, no. 14, 2013.

[68] S. Myers and J. Leskovec, "On the convexity of latent social network inference," 2010.

[69] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," 2011, pp. 561–568.

[70] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, no. 4, 2012.

[71] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," 2013, pp. 23–32.

[72] A. Guille, H. Hacid, C. Favre, and D. Zighed, "Information diffusion in online social networks: A survey," *SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013.

[73] B. Barzel and A.-L. Barabási, "Universality in network dynamics," *Nature Physics*, vol. 9, no. 10, pp. 673–681, 2013.

[74] J. Gleeson, "Binary-state dynamics on complex networks: Pair approximation and beyond," *Physical Review X*, vol. 3, no. 2, 2013.

[75] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier, "Detecting memory and structure in human navigation patterns using markov chain models of varying order," *PLoS ONE*, vol. 9, no. 7, 2014.

[76] P. Singer, D. Helic, A. Hotho, and M. Strohmaier, "Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15, 2015, pp. 1003–1013.

[77] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler, "Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling," *Physical Review E*, vol. 76, 2007.

[78] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers - predicting information cascades in microblogs," in *Proceedings of the 3rd Wonference on Online Social Networks*, ser. WOSN'10, 2010, pp. 3–3.

[79] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda, "Learning diffusion probability based on node attributes in social networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6804 LNAI, pp. 153–162, 2011.

[80] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," 2012, pp. 1145–1152.

[81] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," 2007, pp. 551–556.

[82] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," 2010, pp. 599–608.

[83] F. Wang, H. Wang, and K. Xu, "Diffusive logistic model towards predicting information diffusion in online social networks," 2012, pp. 133–139.