
# Likelihood-based assessment of dynamic networks

Clegg, RG; Parker, B; Rio, M

*IMA Journal of Complex Networks* (2016) Page 1 of 18
doi:10.1093/comnet/xxx000

# Likelihood-based assessment of dynamic networks

RICHARD G. CLEGG*
*Imperial College, London*
*Corresponding author: richard@richardclegg.org

BEN PARKER
*University of Southampton*

AND

MIGUEL RIO
*University College, London*

[Received on 5 October 2016]

This paper deals with the problem of assessing probabilistic models which represent the evolution of a target graph. Such models have long been a topic of interest for a number of networks, especially communications networks. The solution developed in this paper gives a rigorous way to calculate the likelihood of the observed graph evolution having arisen from a wide variety of hypothesised models encompassing many already present in the literature. The framework is shown to recover parameters from artificial data and is tested on real data sets from Facebook and from emails from the company Enron.

*Keywords*: Dynamic networks, network likelihood, network evolution

## 1. Introduction

This paper demonstrates a method for analysing dynamic graph models. It applies the well-founded principles of likelihood-based inference and testing to the graphs to find a rigorous statistical method for comparing graph evolution. The method is demonstrated on both artificial models and on two real datasets, interactions between facebook users (wall posts) and emails between employees within Enron.

Dynamic graph models are a subject of increasing research interest. At the turn of the century it was discovered that several networks appeared to share common properties. These were very different from the properties of the well-studied Erdős–Rényi random graphs [8]. Small-world networks [20] and scale-free networks [3] were observed in a variety of contexts. These three models rely on different stochastic processes to generate graphs, while more recent research explores alternative stochastic processes generating graphs with statistical properties that approximate those exhibited by given networks of interest (e.g. the node/link ratio or degree distribution). These statistics are then compared using a *basket of statistics approach*, which compares a generated graph to the target graph to see if its properties are similar, although generally this similarity is difficult to quantify. We discuss some of these problems in section 2.1.

There is a genuine research need to advance the state of the art by creating a model which produces a rigorous measuring stick which can judge the relative worth of two proposed graph evolution models by comparing them in their ability to explain all stages of a graph's evolution.

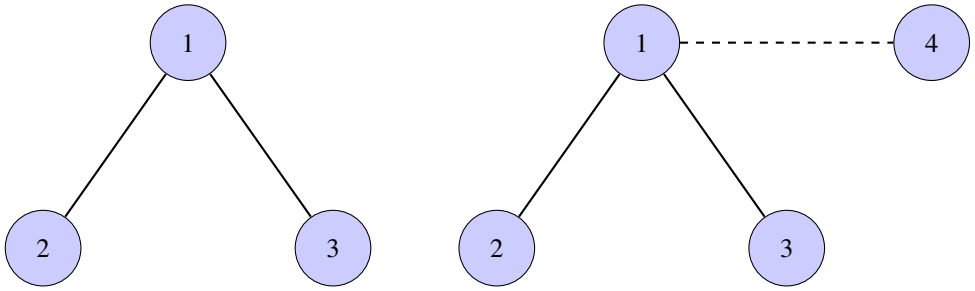


FIG. 1. Example for likelihood based calculation

### 1.1 *Data sets analysed*

In this paper, we consider the analysis of the evolution of two computer-based social networks. Firstly, *facebook* is a dataset consisting of almost 200,000 Facebook wall posts. Secondly, *enron* is a dataset consisting of one million emails sent and received by Enron employees between 1999 and 2003. Both these data sets contain timing information, so we can tell when a connection between two different people is first made. (We assume in this paper that once a connection is made between two people, the connection persists thereafter.) We have chosen these datasets to demonstrate how this method can be useful in social network analysis for networks with different characteristics. In the *facebook* data, posts are made on one person's wall at a time; an e-mail may be sent to many people, so the growth in the *enron* network is quite different.

### 1.2 *Methodology introduced*

This paper suggests a new method for approaching the study of dynamic graphs. When considering rival candidate models for the evolution of graphs it is important to have a reliable measure for which model gives the best explanation for the observed data. Consider the simplified example given in Figure 1 showing only two observerations of a dynamic graph. The graph changes by the addition of a node four and a link between node one and node four. Given these two observations we might compare two hypothetical models for the evolution of the graph.

- Model one. The graph currently containing $N$ nodes evolves by adding one node and an edge connecting it to an existing node randomly chosen with equal probabilities. The probability of choosing node $i$ is $p_i = 1/N$ for all $i$.
- Model two. The graph evolves by adding one node and an edge connecting it to an existing node randomly chosen with probability proportional to node degree. The probability of choosing node $i$ is $d_i / \sum_{j=1^N} d_j$ where $d_i$ is the degree of node $i$.

Here only the node choice section of the model is considered in order to keep the example simple. Model one gives the probability of choosing node one ($p_1$) as 1/3. Model two gives the probability of choosing node one ($p_1$) as 1/2 ($d_1 = 2$, $d_2 = d_3 = 1$). We can see that model two is a more likely explanation for the observed graphs. FETA gives a formal way to calculate the likelihood for a general class of models.

FETA, the *Framework for Evolving Topology Analysis*, is a statistically rigorous approach which can be used to calculate the likelihood that a given sequence of graphs arose from a suggested probabilistic

model. FETA can be shown to recover known parameters from test data where the underlying probabilistic process is known. Further, by directly modelling the graph development processes, FETA can assess probabilistic models which grow graphs with similar characteristics to known target graphs throughout their history. Data and software are available from `https://github.com/richardclegg/FETA2`.

This paper makes several contributions.

- It addresses the problem of determining the process for evolving graphs by presenting a rigorous framework capable of producing a likelihood. Creating a likelihood framework is non-trivial in this context but is necessary if rival models are to be compared rigorously.
- By producing a rigorous likelihood-based explanation of graph evolution, standard techniques in statistical theory become accessible, such as hypothesis testing.
- It accounts for the whole observed lifetime of the evolution of a graph rather than trying to match statistics on snapshots.
- It formally introduces the idea of separating graph evolution models into two parts: the *operation* model which states how the graph will evolve (for example, adding a link, or removing a node) and the *object model* which specifies the probabilities for each link/node. This distinction, detailed in section 3, is implicitly present in earlier work and making it explicit allows the likelihood calculations performed here.
- FETA introduces the concept of breaking evolution models into simpler model components which can be combined proportionally. This allows graph models to be tested.

FETA goes beyond previous contributions by considering the entire graph lifetime, rather than fixating on an arbitrary point in the evolution of the graph.

The aim of the present work is not to describe "the best" models to replicate the evolution of target graphs. Instead, this paper describes a means for determining, given two explanations, which best fits a target graph.

### 1.3 *Structure of this paper*

The structure of the paper is as follows. Section 2 gives further background work. Section 3 describes the FETA framework in detail which also gives some of the assumptions and requirements for data for FETA to be useful. Section 4 describes tests on artificial data. Section 5 describes tests on the model used to select nodes in growing graphs. Section 6 gives conclusions and suggestions for further work.

## 2. Background and related work

Due to the time-varying nature of interactions between network nodes, the modelling of temporal changes in networks has received increased attention from researchers in diverse scientific communities (see, for example, the recent book [12]). However, whereas the temporal analysis of biological networks has focused on the analysis of node clustering changes with time and that of social networks on the changes of network statistics with time [2, 10, 13], communications networks researchers have focused on reachability and graph distance algorithms that explicitly incorporate temporal characteristics [18, 21]. FETA builds upon this previous work and goes further by taking a data-centred approach

to modelling how such networks evolve. It is important to realise from the outset that FETA is not a proposed new model for graph evolution but a method for comparing models of graph evolution.

One of the first models used for graph analysis and generation is the Erdős-Rényi (ER) model, which produces a Poisson degree distribution [8]. Although mathematically convenient, the ER model proved insufficient with the discovery that many networks exhibit *power law* in their degree distributions. The well-known Barabási–Albert (BA) model [3] provided a seminal explanation of these scaling property in terms of a *preferential attachment* mechanism where the probability of connecting to a given node is exactly proportional to its degree. This led to several models which attempted to explain network evolution in terms of node degree and related properties; examples include the [1, 3, 11]. These proved insufficient when a progressively wider set of network statistics was considered, leading to further refinements [5, 15, 25].

It has been shown that a model which faithfully reproduces the node degree distribution may not capture all the important properties of a graph [23]. Different approaches have been proposed. For example, the ORBIS model does not try to model the dynamics of a growing network but, instead, tries to replicate the features of the final network by a process of "rescaling" using graph motifs [14]. In [9] the authors use various spectral measures as summary statistics and approximate Bayesian computation to estimate parameters in graph creation models. All of the papers mentioned here test whether the described model is a good fit to a target graph by considering various graph statistics (see section 2.1). FETA provides an alternative framework for working these models that is both more rigorous and has lower computational demands.

Many networks arising from human communication activity have common structural properties, such as an approximately power law degree distribution. That is the case for the Internet Autonomous System (AS) topology (the network of Internet connections at a large scale), the world wide web hyperlink graph, and social networks based on email exchange and telephone calls (for examples and references see [4, table 3.1]). One common hypothesis for the basis of these shared characteristics is the influence of common network development processes.

In the social sciences, network analysis has shifted from simple models based on descriptive statistics such as node degree or centrality to sophisticated ones that focus on the local selection forces that determine global network structure. In particular, exponential random graph models (ERGMs), also known as $p^*$*-models* [16, 17], can provide optimal estimates for the probability of a new link given the properties of the links and nodes currently in the network. This is achieved by treating the network as a realisation of a set of random variables, and then fitting the parameters for a proposed model for the distribution of these variables. ERGMs can use node properties to predict connections on a static collection of nodes to match a target static graph. This could be seen as similar to what FETA does on dynamic networks. So for a given dynamic model (for example the BA model) for a fixed number of nodes $N$ it is in principle possilble to express the probabilities of all possible size $N$ graphs under BA as an ERGM although the form of the ERGM is not known.

Preliminary versions of this work have been presented at workshops [6, 7]. The former concentrates on applying the results to a number of different data sets, and the latter, to the general theory of FETA and its application to artificial data sets. The work has been considerably developed and clarified since then in terms of terminology and ability to explain real data. This paper further develops the theoretical, modelling and evaluation aspects of FETA, and presents results using two well-known data sets (*facebook* and *enron*). These data sets have been studied before, for example in [19] the authors use the enron data set to investigate automatic detection of change points in graphs. Facebook, of course, has spawned considerable research effort, a summary from the social science perspective is [24].

### 2.1 *The "basket of statistics"*

A typical approach to model graph evolution is to consider a graph at only one time point in its evolution (or a small number of such points) and to measure a number of statistics on it for example [3, 5, 25]. A typical procedure would be to do the following. (1) Take the final state of a network as the "target" for the model. (2) Measure several statistics on this target. (3) Propose a statistical process to generate a graph, for example, "add a new node and connect it to three existing node chosens with probability proportional to its degree" would characterise the preferential attachment model[3]. (4) Use this process to grow a graph to the same size as the target. (5) Measure the same statistics on the artificial graph and compare them to the actual statistics. A model is considered successful if an artificial graph grown according to the model is similar for several chosen statistics. The field has progressed by finding statistics which existing models did not replicate well and showing that improved models replicated those statistics.

Typical statistics would include, number of nodes, number of links, clustering coefficient, the distribution of node degrees, the power law exponent associated with the node degree distribution (if there is one), assortativity, clustering coefficient, graph diameter and many others. Here this approach is referred to as the *basket of statistics* approach. There are a number of problems with this approach. It accounts only for one part of the graph evolution, the end point. The statistics compared may be highly correlated. Comparison of summary statistics may be inconclusive, for example a model is better for the majority of statistics but significantly worse for a minority. The distributions of the statistics are not known, so quantifying how similar one graph is to another is not readily possible. The FETA likelihood framework addresses all these shortcomings.

## 3. A likelihood based framework

The FETA framework produces a rigorous likelihood for a probabilistic model for a graph's evolution. Consider a graph $G(t)$ where links and nodes may be added or removed. Assume that this graph is observed at discrete times (these need not be evenly spaced) to produce a series of graph snapshots $g_i$, where $i \in \mathbb{N}$. Let $\mathbf{g}$ be some sequence of observed graphs $\mathbf{g} = (g_i, g_{i+1}, \ldots, g_{i+n})$; here this paper follows the standard statistical notation that $G_i$ is a random variable and $g_i$ is an observation of it. Define a model $M(\theta)$, where $\theta$ is a vector of model parameters, which determines the probability of all possible next observations given the current and previous observations. The assumed structure of these models will be described in the next section. Let $f_i(g_i|M(\theta))$ be the probability of seeing $g_i$ as observation $i$ given $g_{i-1}, g_{i-2}, \ldots$ for a particular model $M(\theta)$, so that $f_i(g_i|M(\theta)) = \mathbb{P}[G_i = g_i|G_{i-1} = g_{i-1}, G_{i-2} = g_{i-2}, \ldots, M(\theta)]$. (Note that $f_i$ strictly is also conditioned on, and hence a function of, $g_{i-1}, g_{i-2}, \ldots$ but this is omitted for brevity.)

FETA considers the problem of determining the likelihood of the observed evolution sequence $\mathbf{g}$ given a starting graph $g_i$ [1] and a hypothesised explanatory model $M(\theta)$ for the evolution of a graph. It is important to note that at each step of evolution of the model $M(\theta)$ then the probabilities are functions only of the previously observed states of the graph. This allows probabilities to be factored in the normal way. That is

$$\mathbb{P}[G_{i+1} = g_{i+1}, G_i = g_i|G_{i-1} = g_{i-1}, G_{i-2} = g_{i-2}, \ldots, M(\theta)] = f_i(g_i|M(\theta))f_{i-1}(g_{i-1}|M(\theta)).$$

(Note this is need not be a one-step Markovian assumption, $f_i(g_i|M(\theta))$ can be conditioned on the history of observations so far not just $g_{i-1}$.) This can continue for all observations and therefore the

[1] This need not be the earliest observed graph in the sequence.

likelihood of the parameterised model $M(\theta)$ given the observations $\mathbf{G}$ is given by

$$L(M(\theta)|\mathbf{G}) = \mathbb{P}[\mathbf{G}|M(\theta)] = \prod_{k=i+1}^{k=i+n} f_k(g_k). \tag{3.1}$$

The likelihood of the observed sequence is the product of the probability of each step in that sequence.

The question this leaves, of course, is how to produce a probabilistic model which can assign probabilities to different types of changes to a graph. This is achieved by breaking down the model into two parts. These parts are referred to here as the "operation" and "object" models.

- The "operation model" selects the nature of the type of transformation which will be made to the graph; and
- The "object model" elects the exact objects which will participate in the operation.

The literature in the area has implicitly worked with this separation of model elements as will be discussed in the next sections.

### 3.1 *The operation model*

The concepts here will be clearer with examples. In the original Preferential Attachment model [3], the operation model is "At each step, add a new node and select three existing nodes to connect to it." Later work [1] added a more complex model which could, with given probabilities, perform one of three actions to the network, two of which added new nodes and one of which connected and rewired internal nodes. The probabilities could be tuned to replicate the edge/node ratio of the target network. The GLP model [5] uses two operations, the addition of a new node with a fixed number of links and the addition of links between existing nodes. The PFP model in [25] uses a model called the "interactive growth" mode. For the operation model this chooses one from three operations with fixed probabilities. The operations add new links and make onward connections from existing links in different ways.

However, this part of the modelling appears to have received scant attention from previous authors. For example, if the proposed models previously described are used, networks are grown which all generate new nodes and new links in a fixed ratio. This is not what is typically seen in real networks. It seems that if the field of dynamic graph generation models is to progress, one area where progress could be made is in more realistic "operation models". For example, for the enron data described here the link/node ratio varies from near 2.8 at the start of the study period and finishes at near 6.0. For the facebook data the ratio begins at 4.7 and finishes at near 8.0. It is clear that what is termed here the operation model is an important and under-studied problem.

### 3.2 *The object model*

Once an operation model has been specified, the object model works out which entities the operation acts on. In all cases currently used this involves choosing one or more nodes according to a probabilistic rule based upon the current state of the graph. Several such models were mentioned in the introduction. Firstly we define such a model in a common format: The preferential attachment model assigns a probability to a node which which is proportional to its degree. Specifically then, the probability of picking a node $i$ is $p_i = d_i / \sum_{i \in N} d_i$ where $d_i$ is the degree of node $i$ and $N$ is the set of all nodes. All such object models have the common property of requiring normalisation and for simplicity they will

be written in the form $p_i = d_i/k$ where $k$ is understood to be a normalisation constant which makes $p_i$ sum to one over the graph (and which changes with the graph).

An object model then, is a function which makes an assignment between a node in a graph and a probability for choosing that node, in such a way that the probabilities sum to one over all nodes. They will be represented here with $M_\bullet$ and, if the model has a parameter as $M_\bullet(\theta)$. The random model (all nodes equally weighted) will be a null hypothesis and hence referred to as $M_0$. Model components used in this paper are as follows.

- Random model $M_0 : p_i = 1/k$.
- Preferential attachment $M_d : p_i = d_i/k$.
- PFP $M_p(\delta) : p_i = d_i^{1+\delta \log_{10}(d_i)}/k$ where $\delta$ is a parameter.
- Degree power $M_d(\alpha) : p_i = d_i^\alpha/k$ where $\alpha$ is a parameter.
- Triangle model $M_t : p_i = t_i/k$ where $t_i$ is the triangle count of node $i$[2].
- Singleton model $M_1 : p_i = 1/k$ if $d_i = 1,$ or $p_i = 0$ otherwise.
- Doubleton model $M_2 : p_i = 1/k$ if $d_i = 2,$ or $p_i = 0$ otherwise.
- Hot model $M_H : p_i = 1/k$ if node picked in last $n$ picks or $p_i = 0$ otherwise ($n$ is a parameter).

The singleton and doubleton models are used because of the observation that some graphs (notably the AS graph [25] have a different number of nodes of low connectivity than might be expected). The Hot model is introduced because of the observation that in some networks, nodes which have made a connection are more likely to make another connection. Obviously some of these models would become ill-defined at certain points and in those cases the assumption is made that all nodes are equi-probable. For example the singleton model would not otherwise be well-defined if there were no nodes with degree one. The hot model would not be well-defined before any nodes were chosen. By introducing the extra condition of equi-probability then the models are well-defined at any time there is one or more nodes to choose from.

Some of these models well-defined but not useful in isolation (the hot model will pick the same nodes again and again, the singleton model will eventually give every node degree 2). However the models can be linearly combined. So, for example $0.5M_0 + 0.5M_d$ represents a model which is half random and half preferential attachment where $p_i = 0.5/k_0 + 0.5d_i/k_d$ where $k_0$ and $k_d$ are the normalisation constants for the random model and the preferential attachment model respectively. In general models can be constructed by adding together components: for example

$$M(\theta) = \beta_0 M_0 + \beta_p M_p(\delta) + \beta_h M_h(n). \tag{3.2}$$

(random plus degree power plus hot), where $\theta = (\beta_0, \beta_p, \beta_h, \delta, n)$ and this will be a valid probability model if $\beta_\bullet \in [0,1]$ and $\beta_0 + \beta_p + \beta_h = 1$. It is easy to show that if the $p_i$ sum to one (over all nodes in the graph) in each model component then the $p_i$ sum to one if the components are weighted by $\beta_\bullet$ parameters which sum to one. By combining models components in this way a large number of potential models can be generated to test on each graph.

[2]The triangle count is the number of pairs of neighbours of a node that are also themselves neighbours.

Equation 3.1 gives the likelihood of some observed sequence of graph evolution arising from a hypothesised operation model and object model. This can be broken down into the separate likelihood of the operations arising from the operation model and the objects selected arising from the object model. The log likelihood is given by

$$l(M(\theta)|\mathbf{G}) = \log(L(M(\theta)|\mathbf{G})) = \sum_{k=i+1}^{k=i+n} \log(f_k(g_k)).$$

From an observation of a sequence of graphs $\mathbf{g} = (g_i, g_{i+1}, \ldots, g_{i+n})$ and an operation model, a set of object model node choices $C = (c_{i,1}, c_{i,2}, \ldots, c_{i,j(i)}, c_{i+1,1} \ldots, c_{i+n,j(i+n)})$ can be inferred, where at stage $i$ a number of nodes $j(i)$ were chosen, these nodes being $c_{i,1}, c_{i,2}, \ldots, c_{i,j(i)}$. We let the total number of nodes chosen be $N = |C| = \sum_i j(i)$. The likelihood of these objects being chosen can then be calculated using the probabilities $p_i$ for the particular model outlined above. Note depending upon the object model, different numbers of nodes may be chosen at each step.

For the object model, a more human-readable measure for this model is $c_0$, which is is the likelihood ratio of a hypothesised model $M$ compared to the random model under the null model $M_0$.

Definition 1 Let $M(\theta)$ be some object model for the set of node choices $C$ of size $N$, dependent on parameters $\theta$. Let $M_0$ be the random model (all nodes equi-probable). Let $L(M(\theta)|C)$ and $l(M(\theta)|C)$ be the likelihood and log-likelihood of the choices object model $M$ producing choices $C$ with parameter $\theta$. The *per choice likelihood ratio* $c_0$ is the number of observations of $g$ made. It is given by

$$c_0(\theta) = \left[\frac{L(M(\theta))|C}{L(M_0|C)}\right]^{1/N} = \exp\left[\frac{l(M(\theta)|C) - l(M_0|C)}{N}\right].$$

This is introduced for two reasons. Firstly, it enables the experimenter to quickly see whether the proposed object model is better or worse than the hypothesis that the nodes are chosen completely at random, as $c_0 > 1$ represents a more likely model than $M_0$. Secondly, it puts the figures into a more human readable range – for example, in one of the experiments in section 4 the log likelihood varied from $-128868$ to $-118082$ whereas $c_0$ varied from 6.751 to 3.710.

Definition 2 For a particular model $M(\theta)$ parameterised by $\theta$. Let $\Theta$ be the complete parameter space of $\theta$. Given hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta/\Theta_0$, we can form the likelihood ratio test statistic as

$$\Lambda(C) = \frac{\sup\{L(M(\theta) \mid C) : \theta \in \Theta_0\}}{\sup\{L(M(\theta) \mid C) : \theta \in \Theta\}}.$$

In the case where the null hypothesis is nested within the alternative hypothesis, Wilks' theorem [22] tells us that $-2\log\Lambda(C)$ is distributed approximately according to the $\chi^2$ distribution with degrees of freedom equal to the difference in the number of parameters shared between the hypotheses, with the approximation becoming better as $N \to \infty$. This allows us to perform formal hypothesis testing.

EXAMPLE 3.1 Suppose we wish to test whether the model $M(\theta) = \beta_0 M_0 + \beta_d M_d(\delta)$ fits a dataset better than the random model $M_0$ where at each step all nodes are chosen with equal probability. As noted, $\beta_0 + \beta_d = 1$, so we can without loss of generality assume that $\theta = \beta_0$ and parameterise $M(\beta_0) = \beta_0 M_0 + (1 - \beta_0) M_d$. We then wish to test the null hypothesis $H_0 : \beta_0 = 1$ (i.e. that the random model is true) against the null hypothesis $0 \leqslant \beta_0 < 1$. We form the likelihood ratio test statistic $\Lambda(C) = L(M(\hat{\beta}) \mid C)/L(M_0 \mid C)$, where $\hat{\beta}$ is the maximum likelihood estimator for $\beta$, i.e. the value of $\beta$ that yields the choice set C with the highest likelihood. As M contains one parameter, we can test the hypotheses against the $\chi^2$ distribution with one degree of freedom.

Note that when considering a model $M(\theta)$ in which the random model is nested, we can test the hypothesis $H_0$ that the random model is true against the alternative hypothesis that the full model $M(\theta)$ is true for choice set $C$ by considering that $\Lambda(C) = [c_0(\hat{\theta})]^{-N}$ such that the Wilks' ratio $-2\log\Lambda(C) = 2N\log c_0$. This can be tested against $\chi^2_{\nu-1}$, where $\nu$ is the number of parameters in model $M(\theta)$.

### 3.3 *Computational considerations*

To go from the mathematical statements above to computational implementation is a major contribution of this work. The computational implementation of FETA so far (FETA version 2.0) does not make the most general implementation of the framework. In allows analysis of connected simple (no multiple links or self-loops) networks (directed or undirected). The operation model actions comprise the following: add new node and link to $n$ existing nodes and $m$ other new nodes ($n+m>0$); add clique of $n$ new nodes and $m$ existing nodes $n+m>0$; and add a link between two existing nodes.

The code and data and instructions on how to use FETA are freely available at `https://github.com/richardclegg/FETA2`. Work is underway on FETA version 3.0 which will include self-loops, multiple links, weighted links and node and link deletions. It is important to state that these enhancements are intended to reduce limitations of the current codebase, not of the framework.

The workflow for analysing a likelihood for an observed network using FETA 2.0 is described below.

1. Specify an explanatory object model with all parameters – separate models can be included for cliques, new nodes and links between old nodes (see list above).
2. Read the input network which is assumed to be a list of nodes or links each stamped with a time.
3. Where multiple links/nodes are added at the same time:
   (a) Find and remove any sets of cliques added.
   (b) Find and remove any sets of new nodes (and their links) added.
   (c) Anything left is parsed as sets of links between existing nodes.
4. The network is now broken down into a series of operations. The likelihood of each operation is given by the operation model. The likelihood of the objects for the operation is given by the object model.
5. The likelihood with the null object model (random) is also calculated.
6. The likelihoods are combined and output as an overall likelihood, an object model and operation model likelihood and a likelihood ratio per choice relative to null $c_0$.

This gives the likelihood of a proposed model. Often we will want to assess models which are linear combinations of some sub-models, such as in Equation 3.2. In [6] we showed how we might quickly search for the optimal linear combination and hence the model with the maximum likelihood, by regarding the model as a General Linear Model (here without error component) and using standard procedures to find the values of $\theta$ that maximise the likelihood.

Having settled upon one or more models for further investigation the FETA code can be used to output statistics of interest throughout the life of the actual target network and artificial networks generated using the target models.

A number of optimisations are important to make the FETA code run at reasonable speed. For example, nodes are grouped by sets according to their attributes (e.g. sets of nodes of a given degree)

which are tracked as the network evolves. Preferential attachment can then be made faster by selecting a node degree and then choosing one from the selected set of those nodes.

An important note in this is that, when compared with the traditional method of growing example networks and comparing statistics, with FETA it is much faster to assess a likelihood than to grow a network. When assessing a likelihood, at each stage the procedure must calculate the probability of the link and node combination chosen. When growing a network, at each stage the procedure must calculate the probability of every potential link/node that may be picked and then choose one. In the experiments here, assessing the likelihood of a model growing a target network was between 10 and 100 times faster than growing an artificial network using that model – this is another advantage of using likelihoods rather than comparing statistics on grown models.

## 4. Artificial tests

The first test is to see if FETA can estimate known artificial parameters. The results in this section will concentrate on the object model. In this section, the experiments have common form. A target graph is generated from a simple known operation model and object model. The object model will be parameterised. The likelihood of the target graph will then be measured with various settings for the model parameters. If FETA works correctly then the likelihood will be maximised when the parameters match the input parameters used to grow the graph. For these experiments the operation model is very simple. At every iteration a single node is added and connected to two existing nodes. An initial three node network is given (a triangle) and the process continues until 10,000 nodes have been added.

### 4.1 *Plotting likelihood surfaces*

In this section, graphs are grown using object models which have two free parameters and it is shown that FETA can estimate these parameters correctly in most circumstances. Working with the grown network model, every likely pair of parameters is selected and a likelihood surface is produced. Because likelihood detection is much faster computationally than growing a new network then this can be done extremely quickly for modestly sized networks. For example, for the 10,000 link network above, one likelihood test took 0.60 seconds using a single CPU (i7-4710HQ CPU @ 2.50GHz) of a standard laptop. If speed did become a problem then it is not necessary to test every single element of a network's growth to find the maximum likelihood estimator for a parameter. Standard methods in optimisation, such as steepest ascent algorithms, are applicable.

If the FETA procedure provides useful estimates, the likelihood will be highest when the parameters are those which were used to grow the network. In other words, the maximum likelihood estimators found by FETA should be an unbiased estimate for the original parameters for sufficiently large graphs.

The first model tested (model one) is that new nodes are selected with a combination of the degree power model with $\alpha = 0.5$ and the doubleton model. In the terms of section 3.2, the model is $0.5F_2 + 0.5_d(0.5)$. The fitting will assume the form $\beta_1 F_2 + \beta_2 F_d(\alpha)$. Obviously, to be valid $\beta_1 + \beta_2 = 1$ so there are only two free parameters, $\beta_1$ and $\alpha$. The actual values chosen were $\beta_1 = \beta_2 = \alpha = 0.5$. All values of $\beta_1$ from 0.1 to 0.9 were tried (at intervals of 0.05) and values of $\alpha$ from 0.1 to 2.0 were tried (at intervals of 0.1). The likelihood is shown in terms of the ratio $c_0$ defined in Definition 1.

Figure 2(left) shows the likelihood surface for the parameters and contour lines of equal likelihood (these lines are not evenly spaced but are selected to demonstrate how the likelihood increases toward the correct value). As can be seen, the likelihood surface forms a dome with the centre (the maximum

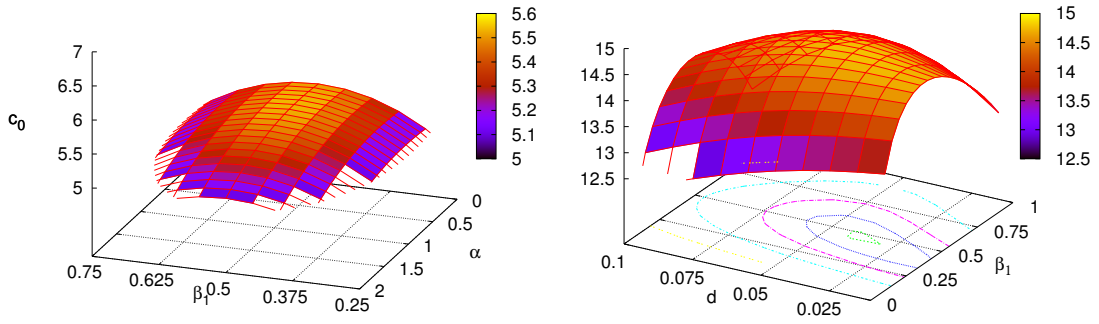


FIG. 2. Artificial network with degree-power/doubleton model (left) and the PFP and triangle model (right) showing $c_0$.

likelihood estimator) in the exactly correct place. From this model, the estimators $\widehat{\beta_1} = \widehat{\alpha} = 0.5$ seem to be unbiased.

A second and more challenging test is provided model two $0.5F_p(0.05) + 0.5F_t$ – that is half PFP with $\delta = 0.05$ and half the triangle model. This is a more difficult test because the triangle model and the PFP model are to some extent doing "the same thing". That is, a node with a high degree is also likely to have a high triangle count. The maximum likelihood estimators are $\widehat{\beta_1} = 0.45$ and $\widehat{\delta} = 0.04$ close to but not exactly equal to the original parameters as can be seen in Figure 2(right).

It is known that maximum likelihood estimators are consistent, in terms of this framework that as the number of graphs in the observed sequence, $n$, increases to $\infty$, the estimator converges in probability to its true value. However, as the graphs analysed are finite it is useful to investigate the properties of the maximum likelihood estimators and specifically that the the bias and variance of the estimators are small for reasonable $n$.

For each of these models, Monte Carlo simulation is performed on graphs grown with Model one and Model two. In each experiment 1,000 graphs are grown for each of Model one and Model two. For each replication the best $c_0$ value was found with $\beta_1$ values tested from 0.35 to 0.65 at intervals of 0.01 and $\alpha$ values tested in the same range and interval. Graphs with 5,000 links and 10,000 links were used to see how the variance in estimates changes with graph size.

A plot of counts for best the $\widehat{\beta_1}$ and $\widehat{\alpha}$ estimates (the values of that maximise $c_0$) for each of the 1,000 runs is shown in Figure 3. As can be seen, both the $\widehat{\beta_1}$ and the $\widehat{\alpha}$ estimates are clustered around the correct values ($\beta_1 = 0.5$ and $\alpha = 0.5$). No significant bias can be seen. For $\widehat{\beta_1}$ the mean values were 0.500025 and 0.5 respectively for 5,000 and 10,000 links. For $\widehat{\alpha}$ the mean values were 0.4993 and 0.498125. the $\beta_1$ parameter was almost always correctly estimated even with the smallest number of counts. The variance in the estimate was low, 0.00237 and 0.000974 respectively for $\widehat{\alpha}$ and 0.0000169 and 0 for $\widehat{\beta}$. This is good evidence that even for relatively small graphs the FETA method produces a low bias and low variance estimator.

The same experiment is tried with model two, again with 1,000 replications and graphs of size 5,000 and 10,000 links. In this case the true values of the parameters are $\beta_1 = 0.5$ and $\delta = 0.05$. The $\beta_1$ values were tested as before from 0.35 to 0.65 at intervals of 0.01 and $d$ values were tested from 0.035

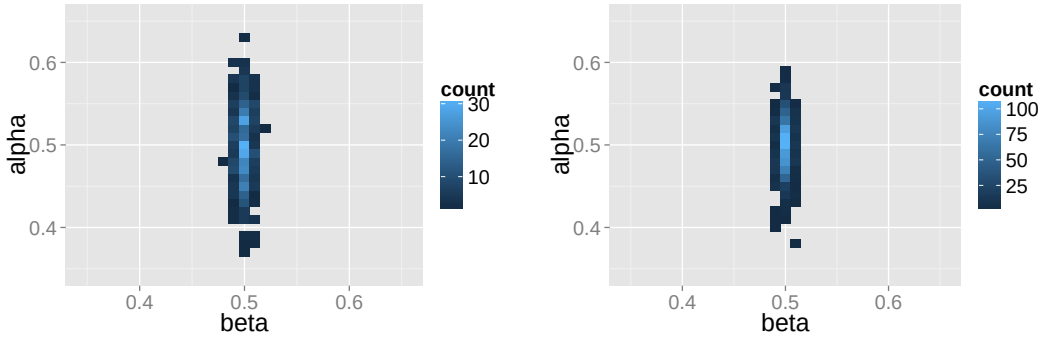


FIG. 3. Counts for best $c_0$ values for model one, 5,000 links (left) 10,000 links (right).

to 0.065 at intervals of 0.001. The resulting counts for the values of $\widehat{\delta}$ and $\widehat{\beta_1}$ (again those that maximise $c_0$) are shown in Figure 4. Again there is good evidence even for relatively small graphs that the bias and variance of the estimate are low and converge quickly. The mean values for $\widehat{\beta_1}$ were 0.499 for 5,000 links and 0.5 for 10,000 links. The mean values for $\widehat{\delta}$ were 0.0496 and 0.0499 respectively. The variances for $\widehat{\beta_1}$ were 0.0000106 and 0.00000304 and for $\widehat{\delta}$ were 0.000253 and 0.000112. Again, this is strong evidence that FETA produces low bias low variance estimates for known parameters. We also see that there appears to be a covariance between $\widehat{\beta_1}$ and $\widehat{\delta}$, reflecting our belief that it is challenging to distinguish between the triangle and the PFP object models.

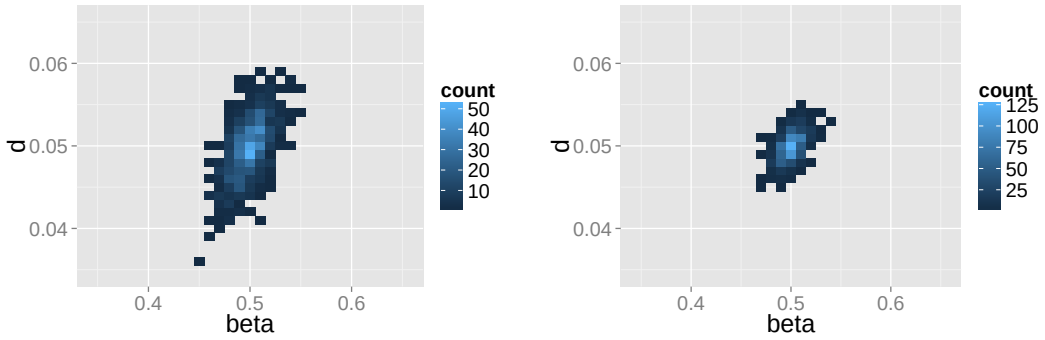


FIG. 4. Counts for best $c_0$ values for model two, 5,000 links (left) 10,000 links (right).

## 5. Object model tests on real data

FETA has already been tested with several data sets [6, 7]. These include an authorship network, two measurements of the Internet Autonomous Systems graph and two photo sharing websites. In addition

to presenting an improved modelling framework, this paper tests it using two additional freely available[3] data sets of interest to communication networks, *facebook* and *enron*.

In this section the focus of interest is the object model. The operation model is isolated by directly using the operations provided by network growth datasets. That is, when in the real dataset a link is added between existing nodes or a new node is linked to one or more nodes, FETA will apply this same operation at exactly the same time. However, instead of choosing the nodes defined by the dataset, FETA instead connects nodes chosen using the object model.

In this section, candidate object models are tested by assessing their likelihood. This testing is done purely by assessing existing data from the target network, not by growing a candidate artificial models and comparing its statistics. The $c_0$ value (higher indicating a better likelihood) is used to distinguish which model is better. If the FETA technique works well, then a higher $c_0$ should lead to an improved match in artificial networks grown using these models.

### 5.1 *Facebook wall post data*

This data set, termed *facebook*, uses facebook wall posts. The network is treated as undirected and duplicate links are removed leaving a network with 183,412 links.

The approach is to compare the real network data with an artificial network grown from probabilistic rules. In this case, the object models compared are random connections, preferential attachment, the PFP model 9with the optimal value of $\delta$ to maximise likelihood), the Degree power model (with the optimal value of $\alpha$ to maximise likelihoods) and a "best" model (chosen by investigating a large number of combined models and selecting the model and parameters with the highest likelihood).

The claim made is not that these models are good representations of the graph but that the likelihood predicts which models are better representations. To reduce the number of plots given on the graphs, only the preferential attachment, the "best" model and the model with the highest $c_0$ from PFP and degree power are shown. In general PFP and degree power had similar success and the random model was, as would be expected, uniformly very bad.

The Preferential Attachment model has $c_0 = 1.091$. The PFP model with $\delta = -0.2$ has $c_0 = 1.199$. The Degree Power model with $\alpha = 0.575$ has $c_0 = 1.220$, The "best" model is only a slight improvement. Mixing the random model with the Degree Power model and adjusting the $\alpha$ to 0.8 improves the likelihood but only to $c_0 = 1.221$. The "best" model is $0.41M_r + 0.39M_d(0.8)$. From these results it would be expected that "best" and Degree Power are approximately as good a fit to the target graph and both are better than preferential attachment.

It is possible to test whether the "best" model is significantly improved on the degree power model. For our general model $M(\beta_0, \alpha) = \beta_0 M_r + (1-\beta_0)M_d(\alpha)$ which is a mixture of the random model plus the degree power model with parameter $\alpha$, we wish to test the null hypothesis that $\beta_0 = 0$ against the alternative hypothesis $\beta_0 > 0$. We form the ratio $\Lambda(C)$ which for this dataset is the ratio of the best model found under the null hypothesis (e.g. that with only the degree power model), where $c_0(M_d) = 1.22025$ against that for the unrestricted model with $c_0(M) = 1.22129$ as $(\Lambda(C))^{1/N} = 1.22025/1.22129$, from which we can calculate the Wilks' Ratio $-2\log\Lambda = -2N\log(1.22025/1.22129) = 216.2$. Comparing this to the $\chi^2$ distribution with 1 degree of freedom suggests that the "best" model is significantly better than that with preferential attachment alone, the p-value being effectively zero.

Having identified three candidate models, PA, degree-power and "best" the next stage is to grow three networks and see how well they fit various statistics during the growth of the graph. To ensure

[3] http://konect.uni-koblenz.de/networks

repeatability each model is grown five times and the statistics averaged. The mean value of the five repetitions are used. The standard deviation was calculated but the error bars were too small to show up on the graphs.

This process of growing a model and comparing statistics with the target graph is typical of the Basket of Statistics approach discussed in the introduction. The flaws of such comparison are as previously discussed. However, it does provide some insight into the connection between graph likelihood and differences in measured statistics across the evolution of the graph.

The plots here are all given as plots of absolute percentage error against time, that is, if at time $t$ the target graph has $S(t)$ for the statistic and the artificial graph has $S'(t)$ then the graphs plot $100|[S'(t) - S(t)]/S(t)|$. While this hides some information (the absolute value of the statistic in question) it provides a much easier comparison between hypothetical graph models. If the answer is closer to zero it is a closer fit for that statistic at that time. All graphs are grown from the same initial seed models and hence start with zero error. Each model is grown five times and averaged, error bars were so small that they did not show up on the plots – the graph models were replicating the statistics very closely between runs.

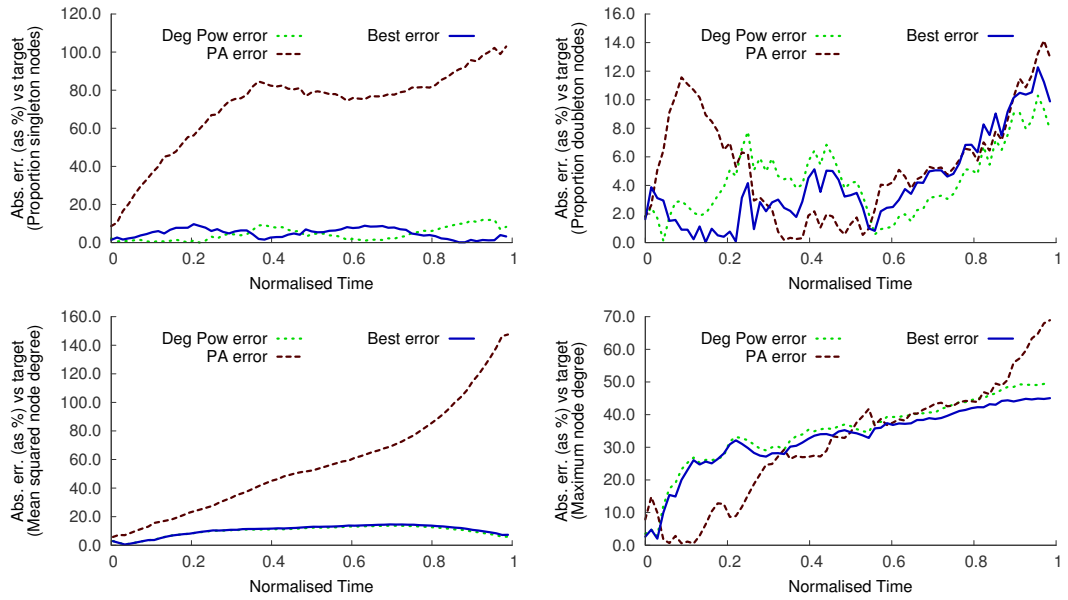


FIG. 5. Error in estimation for the facebook data set: Number of singleton nodes (top left), doubleton nodes (top right), Mean degree squared (bottom left) Maximum node degree (bottom right)

Figure 5 shows the accuracy to which various statistics on the facebook graph are modelled. Four statistics are shown. For the majority it is clear that the degree power and PFP models are fairly realistic. The preferential attachment model performs poorly at estimating the number of singleton nodes and the mean degree squared. Maximum node degree is a notoriously hard statistic to estimate. More importantly, however, the models are as would be predicted from the $c_0$ value, with "best" and degree power performing similarly but both outperforming preferential attachment.

Other statistics showed a similar pattern the exception being clustering coefficient where all three

models performed almost identically and had nearly 100% error (the real data had a CC which began at 0.06 and finished around 0.12 and the models did not predict this growth and uniformly kept the CC of 0.06 throughout). This is likely because none of the models had a triadic closure element (a model component that would preferentially complete triangles). In general though, as would be expected, the $c_0$ value was a good predictor for the success of a model at replicating the real graph.

### 5.2 *Enron data model*

The *enron* data set consists of more than one million emails sent and received by Enron employees between 1999 and 2003. Email recipients/senders are modelled as nodes, and at least one email between them defines a link. The network is treated as undirected and non complex (duplicate emails are ignored). After removal of repeated links, 256,133 links remained. At one point a large email to many people greatly changed the maximum degree (and to a lesser extent other network properties) – this can be seen at time 0.075 on the graph representing maximum node degree (one email to many people becomes the maximum degree node).

A similar process as used for the facebook data was used to find the maximum values for $c_0$. The PFP model has its maximum $c_0 = 4.932$ when $\delta = -0.02$. The degree power model has its maximum $c_0 = 4.907$ with $\alpha = 0.98$. The preferential attachment model gives $c_0 = 4.898$. The similar $c_0$ values indicate that there is little difference between the three models in terms of likelihood (PFP is equivalent to PA when $\delta = 0$ and the degree power model is equivalent when $\alpha = 1$). The "best" model was a combination of the "hot" model (choose again a recently chosen node) and PFP with $\delta = -0.02$ giving the model $0.75F_p(-0.02) + 0.25F_h(1)$. This model has $c_0 = 22.479$ indicating it is considerably more likely than PFP or PA.

The general form of the "best" model is $M(\beta_0, \beta_1, \alpha, n) = \beta_0 M_r + \beta_1 M_h(n) + (1 - \beta_0 - \beta_1) M_d(\alpha)$. If we wish to formally test whether the "best" model has higher likelihood than, for example, the PFP model, we test $H_0 : \beta_0 = \beta_1 = 0$ against $H_1$ : At least one of $\beta_0$ and $\beta_1$ is non-zero. The Wilks' Ratio is -2$\log \Lambda(C) = -2N \log(4.903/22.479) = 518574$ which shows that we can reject $H_0$ with an overwhelming amount of evidence: the "best" model's likelihood is substantially higher.

A similar procedure to that for the facebook data set is followed, and three models are grown five times each (in this case, PFP, PA and "best") and matched to the target data. Again error bars showing the difference between the runs were too small to show up on the graphs. Given the values of $c_0$ it would be expected that PA is worse than PFP which is much worse than "best".

Figure 6 shows the same four statistics for the enron data. This model proved a little harder to match real statistics. In particular the count of doubletons (nodes of degree two) proved hard to model. It is hard to distinguish between the performance of PFP and "best" but it is clear that preferential attachment performs badly on two statistics. Again the clustering coefficient (not pictured) was captured poorly by all models. The graphs shown here are completely consistent with what would be expected from the likelihood with the exception that it would be expected that "best" would show an extremely clear advantage which it does not. The explanation here is almost certainly the hot model component. This strongly predicts the exact order in which links are selected and this does not affect any of the statistics shown. Therefore it is likely that the increased likelihood the "best" model has is demonstrated in factors not shown in the investigated statistics.

In summary then, for both datasets (five others are tested in previous work) the likelihood assessment can be used to test a large number of models extremely quickly. When test networks are grown to compare the models with the real data, models with higher likelihood provide a better fit to observed

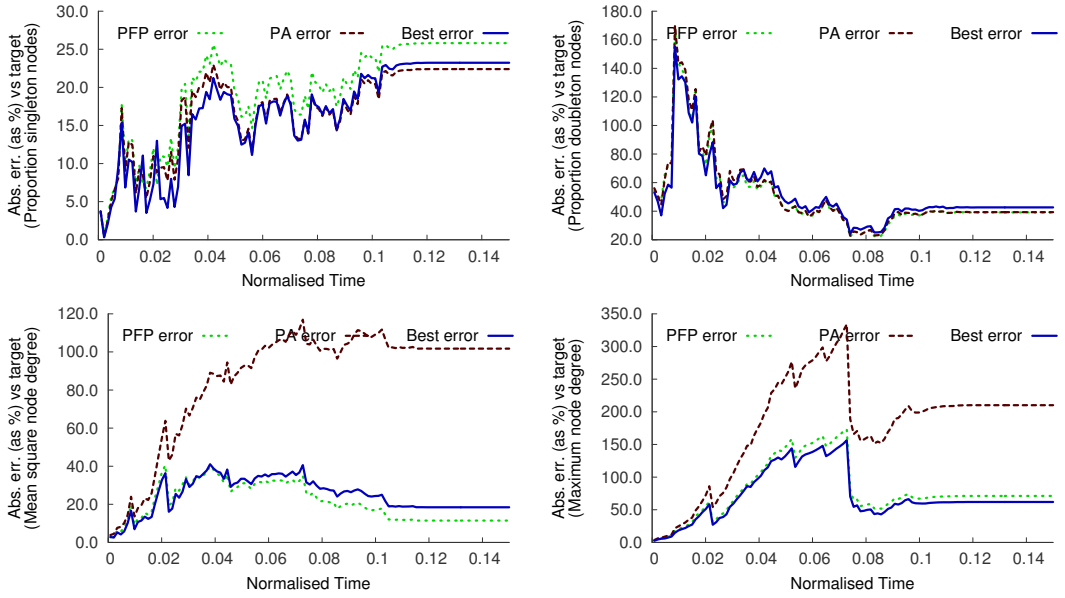


FIG. 6. Error in estimation for enron data: Number of singleton nodes (top left), doubleton nodes (top right), Mean degree squared (bottom left) Maximum node degree (bottom right)

statistics on the graph. Moreover, for nested models, a rigorous test can be used to see if one model is significantly more likely than another.

## 6. Conclusions and further work

This paper introduced FETA (Framework for Evolving Topology Analysis). The framework is used to compare potential models which explain the evolution of growing networks. Previous approaches at modelling graph evolution have attempted to match any of a number of statistics to a particular snapshot of the network at a point in time. Where information is available about the history of a network this throws away the vast majority of the data. FETA, instead, constructs a likelihood for a hypothesised model based on the entire observed section of graph evolution. FETA identifies two model components: an operation model (that selects the nature of the change to the graph) and object model (that selects the actual links and nodes involved with the operation). Using these as separate components in a graph evolution model has previously been implicit in the literature but has never been formalised as such. It is this formalisation that allows the development of a full likelihood model.

The framework is tested on artificial data grown using probabilistic models with known parameters. It is shown that, given the evolution of a target graph, FETA can correctly recover the parameters of the growth model. On real data it is shown that FETA can rapidly test a large number of models to find the best fitting model without having to grow artificial graphs and measure statistics on them. Further it is shown that when graphs are grown, the fit to statistics is in line with the likelihood given by FETA (improved likelihood translates to improved performance on a number of graph statistics).

The operation model is an important part of graph evolution but it has been given scant attention.

Further work on refining what operations the model can use to alter a graph would be an important step and this will form the basis of future work. Graphs where nodes and links can be deleted as well as added could be studied in this way.

FETA is a promising method which looks at probabilistic graph models in terms of the evolution of the graph rather than merely trying to fit a snapshot. The CPU and memory requirements are extremely modest, and an implementation of the framework has been provided.

## REFERENCES


**1.** Albert, R. & Barábasi, A.-L. (2000) Topology of evolving networks: local events and universality. *Physical Review Letters*, **85**, 5234.

**2.** Backstrom, L., Huttenlocher, D., Kleinberg, J. & Lan, X. (2006) Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 44–54, New York, NY, USA. ACM.

**3.** Barabási, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

**4.** Bornholdt, S. & Schuster, H. G., editors (2003) *Handbook of Graphs and Networks*. Wiley.

**5.** Bu, T. & Towsley, D. (2002) On distinguishing between Internet power law topology generators. In *Proceedings of IEEE INFOCOM*, New York, NY.

**6.** Clegg, R. G., Landa, R., Haddadi, H. & Rio, M. (2009a) Measuring the likelihood of models for network evolution. In *Presented at NetSciCom (INFOCOM workshop)*.

**7.** Clegg, R. G., Landa, R., Harder, U. & Rio, M. (2009b) A likelihood based framework for assessing network evolution models tested on real network data. In *Presented at SIMPLEX workshop*.

**8.** Erdős, P. & Rényi, A. (1959) On Random Graphs I. *Publicationes Mathematicae*, **6**, 290–297.

**9.** Fay, D., Moore, A. W., Brown, K., Filosi, M. & Jurman, G. (2014) Graph metrics as summary statistics for Approximate Bayesian Computation with application to network model parameter estimation. *Journal of Complex Networks*, page cnu009.

**10.** Holme, P., Edling, C. & Liljeros, F. (2004) Structure and Time-Evolution of an Internet Dating Community. *Social Networks*, **26**, 155.

**11.** Holme, P., Karlin, J. & Forrest, S. (2008) An integrated model of traffic, geography and economy in the internet. *SIGCOMM Comput. Commun. Rev.*, **38**(3), 5–16.

**12.** Holme, P. & Saramäki, J., editors (2013) *Temporal Networks*. Understanding Complex Systems. Springer.

**13.** Kumar, R., Novak, J. & Tomkins, A. (2006) Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 611–617, New York, NY, USA. ACM.

**14.** Mahadevan, P., Hubble, C., Krioukov, D., ley Huffaker, B. & Vahdat, A. (2007) Orbis: Rescaling Degree Correlations to Generate Annotated Internet Topologies. In *Proceedings of ACM SIGCOMM*.

**15.** Papadopoulos, F., Kitsak, M., Serrano, M. A., Boguna, M. & Krioukov, D. (2012) Popularity versus similarity in growing networks. *Nature*, **489**(7417), 537–540.

**16.** Robins, G., Snijders, T., Wang, P., Handcock, M. & Pattison, P. (2007) Recent developments in exponential random graph ($p^*$) models for social networks. *Social Networks*, **29**(2), 192–215.

**17.** Snijders, T. A. B., Pattison, P. E., Robins, G. L. & Handcock, M. S. (2006) New specifications for exponential random graph models. *Sociological Methodology*, **36**(1), 99–153.

**18.** Tang, J., Musolesi, M., Mascolo, C. & Latora, V. (2010) Characterising temporal distance and reachability in mobile and online social networks. *SIGCOMM Comput. Commun. Rev.*, **40**(1), 118–124.

**19.** Wang, H., Tang, M., Park, Y. & Priebe, C. (2014) Locality Statistics for Anomaly Detection in Time Series of Graphs. *Signal Processing, IEEE Transactions on*, **62**(3), 703–717.

**20.** Watts, D. & Strogatz, S. (1998) Collective dynamics of small-world networks. *Nature*, **393**(6684), 440–442.

**21.** Whitbeck, J., Dias de Amorim, M., Conan, V. & Guillaume, J.-L. (2012) Temporal Reachability Graphs. In *MOBICOM '12: Proceedings of the International Conference on Mobile Computing and Networking*. ACM.

**22.** Wilks, S. S. (1938) The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.

*The Annals of Mathematical Statistics*, **9**, 60–62.

**23.** Willinger, W., Govindan, R., Jamin, S., Paxson, V. & Shenker, S. (2002) Scaling phenomena in the Internet: critically examining criticality. In *Proceedings of the National Academy of Sciences*, volume 99, pages 2573–2580.

**24.** Wilson, R. E., Gosling, S. D. & Graham, L. T. (2012) A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*, **7**(3), 203–220.

**25.** Zhou, S. & Mondragón, R. J. (2004) Accurately modeling the Internet topology. *Phys. Rev. E*, **70**(066108), 1–7.