Research plan

Network structure from group response (NESTOR)

Aristides Gionis

1 Project general information

Project title: Network structure from group response (NESTOR)
Principal investigator: Prof. Aristides Gionis
Contact information: email: `aristides.gionis@aalto.fi`, tel: +358 50 430 1651
Site of research: Department of Information and Computer Science, Aalto University
Duration: 48 months, 1.9.2015-31.8.2019
Scope: 6 person-years of work
Applied funding: 445 134 euros

2 Rationale

Networks (or *graphs*) is a common way to represent entities and their relations, and they are used to model a variety of data. Coupled with the fact that the current technology enables to collect large datasets that can be stored as graphs, and given the rich algorithmic foundations offered by graph theory, network analysis has emerged as a prominent field of data mining.

On the other hand, there are cases where a set of entities can be organized as a network, yet the structure of this network is *unknown* or *unobserved*. There are many reasons for this. Sometimes the network structure is simply *unavailable*. For example, consider a group of analysts blogging on political issues. Many of the analysts may know each other, they may work in the same agency, and they certainly read the posts of each other. A reader of those posts has no way of knowing these connections, although they have a significant role in understanding and interpreting what the analysts are reporting.

In other cases, a view of the network is known but one is interested in *implicit structure* that is impossible to obtain with current data-collection methods. For example, the collaboration network of the employees of a company could be available via information about participation of the employees in past projects. Yet one would like to discover additional information, such as who works well with whom, what are the skills of the employees, who is a good team leader, etc. The previous example with the political analysts, can also motivate the importance of finding implicit structure. Namely, even if we knew the connections among the analysts we would still be interested in understanding who is influencing whom, and on which topic.

This project addresses the problem of *inferring network structure*, explicit or implicit. We assume a set of *entities* for which we want to infer the underlying network structure. We also assume that we have data that record the *observed behavior* of different *groups* (subsets) of entities. We refer to such observed behavior by the term *group response*. The group response contains information regarding which entities are the members of the group, and depending on the application, other information such as description of the context that the group operated, temporal information, performance scores, etc. For an illustration of our concept see Figure 1.

High level objective of NESTOR: Infer the network structure of a set of entities by observing different groups of entities and the response of each group.

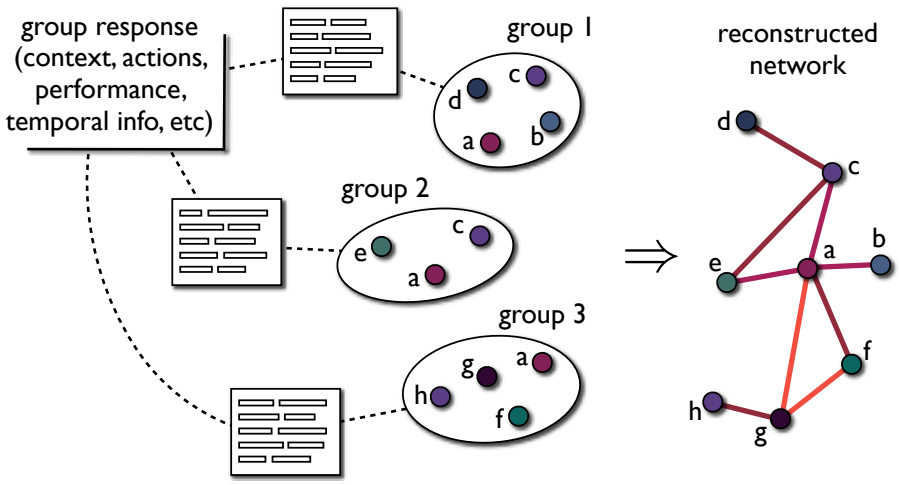

Figure 1: Overview of the project concept. We observe a collection of groups over a ground set of entities $\{a, b, \ldots\}$, and the response behavior of each group. The group response contains information such as context, actions, time, performance, etc. Depending on the application, edges in the reconstructed network may have different semantics, such as influence or synergy, and they may contain additional information, such as probabilities or labels.

Within the general framework of the project we distinguish two cases, which differ with respect to the network structure we aim to infer and the type of group response considered.

Inference of influence networks: In this case, we consider that entities perform actions at different times, and the objective is to estimate whether the actions of one entity influence the actions of another entity and, if so, by what degree. This setting can be used to identify influence between political analysts, in the example given above, or in general to quantify influence between users in social networks. We map this setting to our group response setting, by considering that *the set of entities that have performed the same action defines a group.*

Inference of cooperation networks: In this case we consider that groups entities *collaborate* in order to accomplish different *tasks*, or they form *alliances* that produce different outcomes. The input to the inference problem contains information about the composition of the groups, and depending on the application, additional information such as description of the task or the coalition, observed outcome, or performance score. The objective is to to infer the network structure. In this case, the network edges represent *compatibility* or *synergy* among the entities. We may also be interested in inferring other structure, such as the *role* of the entities in the groups, their expertise, etc.

The problem of inferring influence networks is already an established problem in data mining and machine learning [2, 6, 7, 10, 11, 16, 21]. Our recent work however, suggests that there are interesting new directions to be pursued [4, 22].

On the other hand, the problem of inferring cooperation networks is a new research question that NESTOR will bring forward. Since it is a new problem, we provide some motivating real-world applications.

Collaboration among professionals. Digitization has brought a revolution to communication and collaboration. We now have companies operating with geographically-dispersed teams, open-source initiatives that deliver high-quality products developed by amateurs and enthusiasts, clients outsourcing tasks to freelancers in virtual platforms like `oDesk`,[1]

[1] `http://www.odesk.com`

and so on. Overall, working environments are becoming increasingly more fluid, and it is now a legitimate ambition to expect to be able to match the right person with the right expertise to the right team to deliver the right outcome. To realize this vision, there is need to create tools that support and facilitate collaboration among individuals, and the ability to form efficient teams is essential. NESTOR will contribute to this direction by developing methods for inferring the structure of cooperation networks, which can be used to build skilled and effective teams.

Cooperation among intelligent autonomous agents. Cooperation is encountered not only among humans, but also among intelligent autonomous agents, such as robots or software components. With the upcoming *Internet of Things* revolution, an unprecedented number of devices and systems, covering a variety of protocols and domains, are expected to be interconnected. Those autonomous agents will be expected to cooperate with each other, often in unfamiliar contexts, and in unpredictable configurations. Thus, it is crucial to be able to process the data produced by the existing cooperations of groups of agents and draw conclusions regarding their interaction dynamics, compatibility scores, and synergy potential. A first step to this direction was recently taken by researchers in CMU who addressed the problem of learning *synergy graphs* of autonomous agents [14].

Microbial community characterization. Microbial species form symbiotic communities in different environments, such as different sites of the human body as well as the environment, such as the deep biosphere. Microbial communities are understood to be important players in their respective habitat, yet their structure and function is currently not well understood. Recent advances is DNA sequencing technology has now enabled a more detailed study of microbial communities. Analyzing the data can help us understand which species are working in symbiosis and which are in competition with each other.

In NESTOR we will develop methods to address the network-inference problems discussed above: learning the structure of the influence networks and the structure of cooperation networks. We will explore different variants of the problems of increasing complexity, richness, and ability to model real-world application scenarios. We will study the complexity of the resulting computational problems, and will explore different algorithmic paradigms to develop solutions. We will apply the resulting methods on real-world datasets, which are already available in the scientific community, or we will collect them from the open web and distribute them to the community. The methods developed in the project will be implemented and will become openly available, making the research accessible and reproducible.

Beyond the state of the art. The project will advance the state of the art in the area of social-network analysis. The first thrust of the project (inference of influence networks), follows the recent line of work on the analysis of spread of influence in social networks and modeling the mechanisms of information propagation. The study of this topic has a long history in the social sciences [5, 24], but the area has recently received a lot of attention in the areas data mining and machine learning [6, 7, 10, 12, 21].

The second thrust of the project (inference of cooperation networks), will build on research related to developing algorithms for forming teams of experts. The *team formation* problem aims to assemble a team of experts to accomplish a certain task, by taking into account the expertise and skills of individuals, and their social profile, which includes their social connections and compatibility with others [13]. Previous work, however, assumes that the compatibility network among individuals is given, which, in the majority of applications, is an unjustifiable assumption. With the methods developed in NESTOR we will eliminate the need for this assumption, by *learning* the structure of cooperation networks from past behavior and performance of teams and groups.

We note that the term *network inference* is also used in the scientific literature with different meanings than the one described in this project. The most popular use of the term is in the context of graphical models and learning Bayesian networks [18]. This line of work focuses on learning conditional dependence between random variables, and is often used for modeling casual networks. Our work employs different concepts and has different objectives, although certain techniques from the literature of learning Bayesian networks can be applicable, such as the rule-based learning, EM algorithm, MCMC sampling, etc.

Links of the PI to the research topic. The project builds on the long-standing expertise of the PI on methods for data mining and network analysis. One of the unique strengths of the PI and the team, as demonstrated by past research is to combine successfully theory and practice, by formulating novel problems, designing scalable algorithms with provable quality guarantees, and applying these methods in relevant applications.

The PI has expertise on many of the topics addressed in the proposal. His recent publication record includes work on methods for inferring network structure in propagation networks [4, 15, 22], as well as algorithms for the team-formation problem [1]. The problem of distributing importance scores from groups of entities to individual entities, studied by the PI and his-coauthors [17, 9] is also relevant to the project themes. Additionally, the recent line of work of the PI on discovering dense structures in graphs [8, 20, 19, 23] will enhance the algorithmic toolbox required to solve the problems addressed in the proposal.

The PI leads the Data Mining group in the Information and Computer Science Department of Aalto University. At the moment the group has two postdoctoral researchers, Dr. Nikolaj Tatti and Dr. Michael Mathioudakis, who will contribute in theoretical and application-oriented parts of the project. The expertise of the team of the PI is complemented by world-class expertise in algorithmic theory and machine learning in the department, as well as an extensive network of collaborators within the Helsinki Institute for Information Technology (HIIT). The collaboration and mobility plan (see Sections 8 and 10) involve research visits to reputed research teams in European and US universities. The collaboration plan has been designed to maximize the potential of the project; the hosting research teams have strong interest for the topics of the project, as well as complementary expertise.

Altogether, the team is uniquely positioned to implement this project.

3 Objectives and expected results

The objectives of NESTOR are the following.

O1. Develop a framework for expressing problems of inferring network structure from group response. Identify different classes of network-inference problems depending on the information used to specify group response, which will be motivated by real-world applications. Identify connections within these problem classes, as well as connections with existing problems. Establish complexity results.

O2. Develop efficient algorithms to infer network structure from group response, for the different problem classes defined in **O1**. For the design of algorithmic solutions, combinatorial methods as well as probabilistic-inference methods will be investigated.

O3. Apply the developed methods and perform in-depth analysis on different application domains for influence networks and cooperation networks.

Expected results. The expected output of the project is the development of novel methods that will advance the state of art in the area of network analysis, data mining, and machine learning. The contributions of the project will be aligned with the project objectives, namely:

(*i*) we will provide a new framework to express problems of inferring network structure given groups of entities and information about their behavior; (*ii*) we will identify relevant real-world applications of the proposed framework; (*iii*) we will provide solutions to the suggested problems and will evaluate empirically the developed methods.

NESTOR defines new research directions, and thus, the project is expected to have high impact to the scientific community. Additionally, the project will make possible to obtain networks in cases in which information about groups is available but the network structure is unknown. The benefit in such cases is that once the network structure is known, one can obtain better understanding of the available data by applying the rich toolbox of network analysis.

Critical points for success, alternative implementation strategies. The research questions and application use-cases of NESTOR have been designed so that they have high chances of success, where success is quantified as (*i*) producing a high-quality doctoral dissertation, (*ii*) publishing in first-tier peer-reviewed journals and conferences, and (*iii*) having impact among academic researchers and practitioners. On the other hand, many of our research questions contain high-risk tasks, which are difficult to complete, but whose success will make the research exceptional. The highest risks arise from not being able to provide solutions with theoretical guarantees, or not being able to develop algorithms that will scale to very large data. In such cases, the alternative plan is to experiment with solutions based on heuristics or simplify our assumptions about the formulated problems. To improve scalability, we will investigate methods based on sampling aiming to obtain approximate solutions.

Publication plan, dissemination. We will aim to publish in top-tier forums (Jufo 3 and 2). For publications we will follow the open-access policies suggested by the Academy of Finland. The project datasets and implementations of our methods will become available via the Aalto web-pages and public code repositories (such as `github.com` or `bitbucket.com`). In general we will take all necessary steps to ensure that not only the output of our research is easily reproducible, but also it has large impact and outreach.

To further improve dissemination, the PI has initiated a research blog,[2] where the results of his research group are communicated in a simple manner to a wide audience. Additionally, the PI often gives tutorials in international conferences, and receives invitations to lecture in doctoral schools. Such events will facilitate the dissemination of the results of the project.

4 Research methods and material

Research methods

NESTOR is composed by two foundation modules and one application module.

Foundation module 1 (FM1). Inferring structure of influence networks. Develop methods to infer the structure of influence networks and improve the state of the art in this domain. Following our recent work [22], and as it will be discussed shortly, the emphasis of our approach will be to use *context* information.

Foundation module 2 (FM2). Inferring structure of cooperation networks. Develop a framework for expressing problems of inferring cooperation network structure from group response. Identify interesting problem formulations, motivated by real-world applications, establish the complexity of those problems, and develop efficient methods.

Application module (AM). The methods developed in the foundation modules will be tested on real-world applications. The datasets used for experimental evaluation will come from

[2] `https://blogs.aalto.fi/data/`

the domains of *social media*, *bibliographic databases*, *project management data*, and *microbial community data*.

Before discussing in detail the work modules we will provide some remarks regarding our methodology.

– **Combinatorial optimization vs. probabilistic inference methods.** The PI has expertise on developing data-analysis methods based on combinatorial-optimization techniques: formulate the data-mining problem as an optimization problem and design combinatorial algorithms to provide exact or approximate solutions. Recently, building on collaboration with professor Rousu, the PI has explored methods for learning network structure with probabilistic-inference and kernel methods [22]. Carrying on the collaboration, we will investigate and compare solutions following both paradigms.

– **Parsimony.** Principles based on parsimony, such as MDL, play an important role in data mining for learning models that avoid over-fitting. In our previous work on network inference, we have used the parsimony principle to formulate problems of finding *sparse networks* [4, 15]. The aim has been to identify the *backbone* of a network. In addition to reducing the noise, sparse structures offer advantages in visualizing and interpreting more easily the available data. In NESTOR we will employ the parsimony principle with a similar rationale. For instance, one way to formulate a class of network-inference problems is to ask to *find a network with the minimum number of edges that satisfies constraints provided by certain group-response input*.

We now provide more details about the work modules of NESTOR.

FM1: Inferring structure of influence networks. The objective of this work module is to infer structure in influence networks. In this case, a *group* consists of the set of nodes that have performed the same action. For example, in a social-media application, actions represent the messages posted by users, while in a scientific-network scenario actions represent papers written by scientists. In both of these examples, additional information is provided for each group, such as, keywords that describe the actions, or time-stamps that describe when each person performed each action. The network-inference problem typically asks to find a set of edges $\{(u, v)\}$, and associated probabilities $p(u, v)$, indicating that if node u performs an action then with probability $p(u, v)$ node v will perform the same action. In some cases the graph structure is given, and the research question is to estimate the edge probabilities. As discussed previously, the problem of network inference in this setting has already received considerable attention [2, 6, 7, 10, 11, 16, 21].

In our previous work we studied new formulations of this problem based on parsimony, thus, asking to infer sparse networks [4, 15]. More recently we exploited information associated with the *context* of the actions [22]. The main idea is motivated by the observation that the influence between two nodes in the network does not depend only on the nodes and their connections, but it also depends on the action under consideration. For example, a person v may be influenced by a person u regarding certain topics, say science, not not regarding other topics, say politics.

The idea of using a *context-sensitive* model was proven quite powerful, and we were able to outperform state-of-the-art methods. On the other hand, the proposed method uses a maximum-margin structured-learning approach and it scales only to medium-size networks. Thus, one of the main challenges is to improve the methods for learning context-sensitive influence networks and make them applicable to very large networks.

One promising approach to take context into account and yet obtain scalable algorithms, is to formulate the network-inference problem in connection with obtaining a *clustering* of the actions that have been propagated in the network. The idea is to cluster the set of actions so that each cluster corresponds to a *topical community*, and within each community the inferred network structure matches well the observed data. We will thus need to develop similarity measures between action propagations, and corresponding clustering algorithms.

Since in many cases the groups associated to action propagations correspond to directed acyclic graphs (dags), a simplified version of this problem is to develop methods for clustering dags. Another reasonable simplification of the problem is to assume that the actions form a *hierarchy*, and the potential action clusters are given only by internal nodes in this hierarchy.

Finally, we will consider inferring network structure not only in terms of network edges, but also in terms of *node labels*, corresponding to the different *roles* that nodes may have in the information-diffusion process. The assumption here is that different topical communities are likely to contain nodes that are expected to assume certain roles: influential nodes, early-adopters, grass-root followers, and so on, where those roles are expected to be encountered with respect to the community-relevant topics and with respect to the network dynamics. We will thus formulate models that capture this intuition and we will develop methods to learn such models and associated network structure.

FM2: Inferring structure of cooperation networks. To better discuss the case of inferring structure of cooperation networks, let us consider the *team-formation problem* [13]: consider a social network consisting of individuals who have certain skills. Given a task, represented as a set of skills, the goal is to find a team in the social network, whose members have the skills required to accomplish the task and who can communicate efficiently. Assuming that the network edges are associated with weights that express the *communication cost* of individuals in the network (where communication is inversely related to *synergy*), the goal is to find a team with as small cost as possible. Lappas et al. [13] suggest some ways to express the cost of a team in relation to the cost of network edges. In particular they consider as team cost the *minimum spanning tree* cost and the *diameter* of the graph induced by the team.

In the setting described above, the problem of inferring structure of cooperation networks can be expressed as follows: given a set of past projects, groups of individuals who have worked on those projects, and other related information, such as skills required for each project, and performance of the groups, can we learn the structure of the underlying cooperation network?

To approach successfully this network-inference problem we will consider different variants of increasing richness and complexity. Our problem variants will be categorized along three different dimensions: (*i*) the underlying model for expressing group synergy; (*ii*) the type of information given as input (group response); and (*iii*) the type of network structure that we will aim to learn.

With respect to the underlying model for expressing group synergy, in addition to the *minimum spanning tree* and *diameter*, considered by Lappas et al. [13], we will consider simpler functions, such as *connectivity*, *sum of edge weights*, and *star constraints* (diameter ≤ 2), as well as more complex functions, such as the *edge density* of the subgraph induced by the group.

With respect group-response information, we will study different cases, depending on whether we are given (*ii*.1) only the members of each group; (*ii*.2) additional keywords describing the task/conditions that each group is involved; (*ii*.3) performance scores for each group.

Finally, with respect to learning different types of network structures, we will consider the following cases: (*iii*.1) learning only the edges of the network; (*iii*.2) learning synergy scores for the network edges; (*iii*.3) learning additional information for the network nodes, such as, expertise score on different topics, or social roles, such as who is a good group leader.

As an example of a problem that one can formulate within the framework described above, consider the following problem (in fact, a special case of the problem we formulated in our discussion about *parsimony*):

Problem 1 (Minimum edge group connectivity) Given a set of entities V and subsets $S_1, \ldots, S_m \subseteq V$ find a set of edges E so that in the graph $G = (V, E)$ all sets S_i *induce a connected subgraph*, and the number of edges $|E|$ is *minimized*.

This problem has been studied by Angluin et al. [3], and it was showed to be **NP**-hard, while the objective function is *submodular*, yielding a $\mathcal{O}(\log n)$ approximation.

We are currently studying different formulations of the problem of learning the structure of cooperation networks, such as, assuming that the group synergy function is the sum of edge weights, or assuming that the group should form a star structure.

Finally, one interesting research direction is to consider an *adaptive* setting and learn network structure in a *explore-exploit* fashion. Namely, assume that we not only observe previous group responses, but we have the opportunity to create teams to perform new tasks. Then, given a task, how should we form a team that is expected to perform well for the task, and at the same time its observed response maximizes our knowledge for the network structure?

AM: Application module. All developed methods will be evaluated on real-world datasets. We will test the validity of our problem definitions by the means of cross-validation, namely, we will use the reconstructed networks to predict response for previously unseen groups. We will compare the solutions found by competing methods in terms of accuracy and we will test the scalability of our methods on large datasets.

Methods developed in **FM1** will be tested on datasets from the domains of social media and bibliographic databases. In the first case, the underlying network structure expresses influence between users posting content in social media, and in the latter case influence between scientists working on different research areas.

The methods developed in **FM2** will again be evaluated on bibliographic data. In this case, instead of considering as group response the set of researchers who have worked on a particular topic and trying to infer the influence patterns, we will consider as a group the set of co-authors of a paper. Meta-information about the publication will be used as context, and the number of citations of the publication will be used as proxy for the performance of the group. In this application, group performance does not depend only on the overall group synergy but also on the expertise and the skills of each individual member, thus it is important to experiment with methods that reconstruct this information, as well. Additionally, we will experiment with project management data, where individuals participate in different projects. A relevant dataset can be obtained from `github`.

Finally, we will experiment with microbial community data, where the groups consist of microbial species and the context is provided by either human intestinal tract or deep bedrock groundwater. The interpretation of the results will be carried in collaboration with professor Rousu and biology collaborators.

Research material

Datasets. Whenever possible we will use publicly available datasets, in order to ensure reproducibility of our results. For the applications in the social-media domain, many datasets are publicly available,[3] while we will also collect `twitter` data via the public API. For bibliographic data, we will use the data provided by DBLP,[4] as well as the Thomson Reuters ISI Web of Science dataset, available to us via the Aalto University license. For the project management use case we will use publicly available datasets from `github`.[5] Additionally, we are currently discussing with a research team in `oDesk` the possibility of obtaining a sample of freelancer project data from them. Finally, publicly available datasets on microbial communities can be obtained from the data sources of the Human Microbiome Project (HMP).[6] In addition, we have access to deep bedrock microbial community data through the collaborators of professor Rousu.

[3] For instance in SNAP: `http://snap.stanford.edu`

[4] `http://dblp.uni-trier.de`

[5] `http://githubarchive.org`

[6] `http://www.hmpdacc.org/resources/data_browser.php`

5 Support from research environment

The host institute for NESTOR is the department of Information and Computer Science of Aalto University. With a wide range of computing resources and support services, a truly international community, and its commitment to high-quality research and teaching, Aalto University provides an excellent supporting environment for NESTOR. The department of Information and Computer Science has long-standing tradition in the areas of machine learning and data mining, and it attracts a continuous stream of bright PhD students and postdoctoral researchers from Finland and abroad.

Additionally the PI serves as the director of the Algorithmic Data Analysis (ADA) programme of the Helsinki Institute for Information Technology (HIIT), a renowned institute for basic and applied research on information technology. HIIT's current foci of research on computational modeling and data analysis will provide further support for NESTOR and will enhance its chances for success.

6 Ethical issues

The ethical issues of NESTOR are related to accessing information of individuals from social-media content. In most cases, the datasets that will be used for experimentation are already publicly available and properly anonymized. In addition, datasets collected within the project will be assembled via calls to public APIs (thus the data are again public) or will be accessed under license agreements. In all cases, our methods will handle data in aggregate forms, and no information about a particular individual will ever be isolated.

7 Implementation: schedule, budget, distribution of work

Timetable and distribution of work. The project is organized in two foundation modules, **FM1** and **FM2**, and one application module, **AM**. The first foundation module will extend during the first half of the project. The second foundation module has broader scope and involves many new problems, thus it will extend during the whole duration of the project. The application module will start after the first six months, so that theoretical concepts have been introduced. However, it is important to start experimenting with applications early on, so that findings from real-world applications can give feedback for theoretical development.

The project is estimated to require the work of one full-time postdoctoral researcher NN1 (2 person years, funding applied from Academy of Finland) and one doctoral student NN2 (4 person years, funding from the Academy). Additionally, the current postdoctoral researchers of the team, Dr. Mathioudakis and Dr. Tatti, will contribute to the project. NN1 will contribute on **FM1** and on building the foundations for **FM2**. NN2 who will be trained in research will contribute on **FM2** and **AM**. In particular, **FM2** will provide the basis for the PhD thesis of NN2. The PI will contribute in all modules via supervision of the student and the postdocs and scientific lead.

All researchers involved in the project will travel to international computer-science conferences to present the project publications. In addition, the PI will travel for research visits within the scope of the project (see collaboration, Section 8, and mobility plan, Section 10). One similar research visit has been planned for the postdoc NN1. The doctoral student NN2 will be encouraged to do summer internships, but funding will be sought elsewhere, e.g., from the HICT doctoral education network. A detailed illustration of the NESTOR time schedule, including the work modules and the mobility travel is shown in Figure 2.

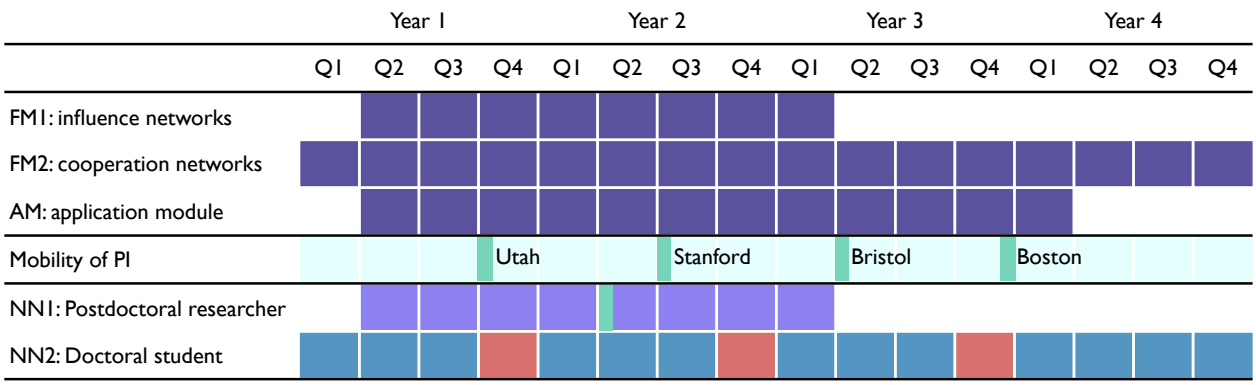

Figure 2: The NESTOR time schedule

Table 1: The NESTOR budget

	2015	2016	2017	2018	2019	Total
Salaries						
Doctoral student	3.5 months × 2 833	11 months × 2 917	11 months × 3 005	11 months × 3 095	7.5 months × 3 189	**133 021**
Postdoctoral researcher		11 months × 3 713	11 months × 3 825			**82 918**
Salaries, total	9 916	72 930	75 130	34 045	23 918	**215 939**
Indirect costs, total	5 255	38 653	39 819	18 044	12 677	**114 448**
Total overheads share	10 620	78 108	80 464	36 462	25 617	**231 271**
Other costs						
Travel expenses						
(conferences)		10 000	10 000	10 000	10 000	**40 000**
(mobility)		5 100	3 050	3 050	3 050	**14 250**
Publication costs						
(open-access fees)		5 000	5 000	5 000	5 000	**20 000**
Other costs, total		20 100	18 050	18 050	18 050	**74 250**
Total costs	25 791	209 791	213 463	106 601	80 262	**635 908**
Funding from own organization						
	7 738	62 938	64 039	31 980	24 079	**190 774**
Funding applied from the Academy						
	18 053	146 853	149 424	74 621	56 183	**445 134**
Academy contribution %	70	70	70	70	70	70

Budget. The largest fraction of the budget is allocated to the salaries of NN1 and NN2. Each of the research visits of the PI and NN1 is planned for one month, and it has been budgeted for 3 000 euros (1 050 travel allowance plus expenses). Additional budget has been allocated to all project members for traveling to conferences. In addition, 5 000 euros per year have been budgeted for open-access publication fees. The budget details are provided in Table 1.

8 Research team, collaboration

Researchers. The PI, Aristides Gionis, was appointed an associate professor at the Department of Information and Computer Science in Aalto University in 2013. Professor Gionis' research focuses on developing data-analysis algorithms for problems motivated by real-world applications. His expertise spans a range of different areas, such as graph mining, web mining, sequence analysis, similarity search, streaming computation, and privacy preservation. According to Microsoft Academic ranking, he is among the top data-mining researchers worldwide. According to Google scholar, his papers have been cited more than 5000 times since 2009, and

eighteen of his publications have received more than 100 citations.

International collaborators. The PI has an extensive collaboration network. Overall he has publications with more than 90 coauthors, coming from all over the world. One currently active channel of collaboration involves researchers from Yahoo! Labs, and the PI recently received the Yahoo! Faculty Research and Engagement Program (FREP) grant. Within the scope of NESTOR, ties with existing collaborators will be enhanced, and opportunities for new collaborations will be sought. The key players are the following.

Professor **Suresh Venkatasubramanian** from the **University of Utah**, is a leading researcher in algorithms for computational geometry, with current focus on data mining and large-scale data analysis. He is currently working on problems of network analysis and from preliminary discussion he is quite interested in the project themes. He will contribute on the theoretical questions of NESTOR and on issues regarding approximations and scalability.

Professor **Jure Leskovec** from **Stanford University**. Professor Leskovec is one of the most prominent and well-cited researchers in data mining. His work has been very influential and he has pioneered many new research directions. His work includes a number of important papers on network inference, and thus he is expected to have significant contribution to the project.

Professor **Tijl De Bie** from the **University of Bristol**, is also one of the most prominent researchers in data mining, and he recently received an ERC consolidator grant. His expertise on maximum-entropy methods can provide strong support for dealing with uncertainty in the project research questions.

Professor **Evimaria Terzi** from **Boston University** is another top researchers in data mining. Her contributions include a number of important and highly-cited publications, including the work that introduced the team-formation problem [13]. Professor Terzi is a long-term collaborator of the PI.

9 Researcher training and research careers

The PI has formed a research group in Aalto University with funds from his tenure-track startup package, HIIT, and the HICT doctoral education network. The group is active, and although the doctoral students are just starting their degree, they have already obtained publications in top-level forums. As an example, Polina Rozenshtein received the best student paper award in ECML PKDD 2014, with a paper based on her Master thesis, while Polina now continues her research as a doctoral student.

The project will hire one PhD student and one postdoctoral researcher via open calls. The education of the PhD student will be closely monitored and supervised, but the student will also be given the required intellectual freedom to develop their own ideas. The PhD student will have the opportunity to work with the postdocs of the group, to travel to conferences, and to make internships in other institutes and industrial research labs. Researcher careers are advanced by the international collaboration network of the project and the group, which will create important contacts to academia and the industry.

In the hiring process, among equally qualified candidates, female applicants will be favored.

10 Mobility plan

The mobility plan of NESTOR consists of four one-month visits, in the research teams mentioned in Section 8: Utah, Stanford, Bristol, and Boston. This is essential for acquiring new knowledge and getting inspiration for the project. The order of the visits was selected so that visits to the groups with whom there is currently no active collaboration (Utah and Stanford) come first, so

that the communication channel will open during the whole course of the project. The teams with which there is already ongoing collaboration or it is easier to pursue one (strong ties with Terzi and often meetings with De Bie in European conferences and workshops) will be visited in the second half of the project, even though discussions will start earlier.

References

[1] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *WWW*, 2012.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. *KDD*, 2008.

[3] D. Angluin, J. Aspnes, and L. Reyzin. Network construction with subgraph connectivity constraints. *Journal of Combinatorial Optimization*, 2010.

[4] F. Bonchi, G. D. F. Morales, A. Gionis, and A. Ukkonen. Activity preserving graph simplification. *Data Mining and Knowledge Discovery*, 27(3), 2013.

[5] J. Coleman, H. Menzel, and E. Katz. *Medical Innovations: A Diffusion Study*. Bobbs Merrill, 1966.

[6] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schoelkopf. Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm. In *ICML*, 2014.

[7] N. Du, Y. Liang, M.-F. Balcan, and L. Song. Influence Function Learning in Information Diffusion Networks. *ICML*, 2014.

[8] E. Galbrun, A. Gionis, and N. Tatti. Overlapping community detection in labeled graphs. *Data Mining and Knowledge Discovery*, 28(5-6), 2014.

[9] A. Gionis, T. Lappas, and E. Terzi. Estimating importance via counting set covers. In *KDD*, 2012.

[10] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *TKDD*, 5(4), 2012.

[11] M. Gomez-Rodriguez, J. Leskovec, and B. Scholkopf. Structure and dynamics of information pathways in online media. In *WSDM*, 2013.

[12] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[13] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, 2009.

[14] S. Liemhetcharat and M. Veloso. Weighted synergy graphs for effective team formation with heterogeneous ad hoc agents. *Artificial Intelligence*, 208, 2014.

[15] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. Sparsification of influence networks. In *KDD*, 2011.

[16] S. Myers and J. Leskovec. On the Convexity of Latent Social Network Inference. In *NIPS*, 2010.

[17] P. Papapetrou, A. Gionis, and H. Mannila. A Shapley value approach for influence attribution. In *PKDD*, 2011.

[18] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[19] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *KDD*, 2014.

[20] P. Rozenshtein, N. Tatti, and A. Gionis. Discovering dynamic communities in interaction networks. In *ECML PKDD*, 2014.

[21] K. Saito, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. In *KES*, 2008.

[22] H. Su, A. Gionis, and J. Rousu. Structured prediction of network response. In *ICML*, 2014.

[23] C. E. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. A. Tsiarli. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *KDD*, 2013.

[24] T. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, 1955.