16

Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

MMB & DFT 2014
Proceedings of the International Workshops SOCNET 2014 and FGENET 2014

Kai Fischbach, Marcel Großmann,
Udo R. Krieger, Thorsten Staake (eds.)

TAO
Technologie
Allianz
Oberfranken

University
of Bamberg
Press

16 Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Band 16

University
of Bamberg
Press
2014

MMB & DFT 2014
Proceedings of the International Workshops

Modeling, Analysis and Management of Social Networks and their Applications (SOCNET 2014)
&
Demand Modeling and Quantitative Analysis of Future Generation Energy Networks and Energy-Efficient Systems (FGENET 2014)

Kai Fischbach, Marcel Großmann,
Udo R. Krieger, Thorsten Staake (eds.)

University
of Bamberg
Press
2014

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.ddb.de/ abrufbar

Herstellung und Druck: Digital Print Group, Nürnberg
Umschlaggestaltung: University of Bamberg Press, Andra Brandhofer

http://www.uni-bamberg.de/ubp/

ISSN: 1867-7401
ISBN: 978-3-86309-208-5 (Druckausgabe)
eISBN: 978-3-86309-209-2 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-64867

Contents

Organization

The MMB & DFT 2014 International Workshops were organized by Faculty of Information Systems and Applied Computer Sciences, Professorship of Computer Science, on behalf of University of Bamberg, Germany.

Organizing Committee

General Chairs

SOCNET 2014

Kai Fischbach	University of Bamberg, Germany
Udo R. Krieger	University of Bamberg, Germany

FGENET 2014

Udo R. Krieger	University of Bamberg, Germany
Thorsten Staake	University of Bamberg, Germany

Local Arrangement Co-Chairs

SOCNET 2014

Oliver Posegga	University of Bamberg, Germany
Cornelia Schecher	University of Bamberg, Germany

FGENET 2014

Marcel Großmann	University of Bamberg, Germany
Cornelia Schecher	University of Bamberg, Germany

Organisational Support

Technical Program Committee

SOCNET 2014

Jana Diesner	University of Illinois at Urbana-Champaign, USA
Kai Fischbach	University of Bamberg, Germany
Peter A. Gloor	Sloan School of Management, MIT, USA
Udo R. Krieger	University of Bamberg, Germany
Katharina A. Zweig	TU Kaiserslautern, Germany

Additional Reviewers

Marcel Großmann	University of Bamberg, Germany
Darko Obradovic	TU Kaiserslautern, Germany
Matthäus Paul Zylka	University of Bamberg, Germany

FGENET 2014

Hermann de Meer	University of Passau, Germany
Reinhard German	University of Erlangen-Nuremberg, Germany
Udo R. Krieger	University of Bamberg, Germany
Paul Kühn	University of Stuttgart, Germany
Michael Menth	University of Tübingen, Germany
Johannes Riedl	Siemens AG, Germany
Thorsten Staake	University of Bamberg, Germany

Additional Reviewers

Ilya Kozlovskiy	University of Bamberg, Germany
Gergoe Lovasz	University of Passau, Germany
Mariya Sodenkamp	University of Bamberg, Germany

Preface

Today, the global increase in the energy demands arising from growing mobility, the industrial and private emissions and the Internet traffic, as well as the steadily increasing depletion of fossil resources generate some of the greatest challenges for engineering and science in modern information societies. The dissemination of multimedia information that is caused by underlying, massive human interactions and supported by a variety of online social networks constitutes one major driver of this fast-growing Internet traffic.

The term social network denotes the social structure that emerges from human actors' interaction among each other. Over the years, scholars in the fields of anthropology, sociology, psychology, economics and organizational theory have proposed different methods and techniques for discovering these structures and drawing conclusions about the functioning of social networks and network outcomes. Practical applications of social network analysis constitute a very important, rapidly growing area in modern interconnected information societies. The related scientific concepts incorporate a variety of methods from areas like graph theory, computer science, and statistics, among others.

If we consider the common mathematical foundation of currently applied modeling, network analysis and performance evaluation techniques, these developed engineering methodologies can also be applied to the design of effective next-generation energy networks and energy-efficient systems.

For this reason, the Program Committee of the 17th International GI/ITG Conference on "Measurement, Modelling and Evaluation of Computing Systems" and "Dependability and Fault-Tolerance" (MMB & DFT 2014), held during March 17–19, 2014, at University of Bamberg in Germany, organized two satellite workshops on March 19, 2014, that covered corresponding research topics:

- an International Workshop on Demand Modeling and Quantitative Analysis of Future Generation Energy Networks and Energy-Efficient Systems (FGENET 2014)

  FGENET 2014 focused on modeling, analysis and simulation of future generation energy networks based on renewable energy sources, attached energy-efficient systems, demand monitoring and planning methods. It pointed out how engineering methodologies that were developed over the last decades to cope with performance, dependability and reliability analysis of general networks help us to understand retail energy markets, to manage energy transition towards more sustainable systems, to design future smart energy networks with storage grids and integrated self-monitoring, intelligent communication and control systems, and to effectively manage the energy-performance trade-off in attached data centers.

- an International Workshop on Modeling, Analysis and Management of Social Networks and their Applications (SOCNET 2014)

  SOCNET 2014 was interdisciplinary in nature, attracting contributions from different fields, as varied as information systems, information science, business administration, computational social science, and computer science. It fulfilled the objective to fully incorporate the rich methodological, technical, socio-economic as well as psychological aspects of social network analysis. Regarding the structural inference in online social networks or the matching of peers with common features, the detection of communities within networks, the mathematical formulation and analysis of dissemination processes in directed social networks, and the visualization of network dynamics, the contributions ranged from mainly theoretical achievements and algorithmic improvements to practical applications such as detection algorithms for protein networks or the performance of distributed problem solving in social groups.

After a careful review and selection process, the MMB & DFT 2014 International Workshops' Program Committees finally generated a scientific program that included 13 regular papers and three additional invited talks:

1. "Network Analysis Literacy", presented by Professor Dr. Katharina Anna Zweig, TU Kaiserslautern, Germany

2. "Coolhunting for "Honest Signals of Innovation" in Social Media", presented by Dr. Peter A. Gloor, Center for Collective Intelligence, Massachusetts Institute of Technology, USA

3. "Modelling of the Worldwide Electricity Consumption of ICT", provided by Ward Van Heddeghem on behalf of Prof. Dr. Mario Pickavet, iMinds-IBCN, Ghent University, Belgium.

We thank all the authors for their submitted papers and all the speakers, in particular the invited speakers, for their lively presentations.

As conference chairs, we are grateful for the support of all members of the Program Committees and thank all the external reviewers for their dedicated service and for the timely provision of their reviews.

We express our gratitude to University of Bamberg as conference host and Technology Alliance Oberfranken (TAO), as well as to all the members of the MMB & DFT 2014 International Workshops' local organization committees for their great efforts.

We acknowledge the support of the EasyChair conference system and express our gratitude to its management team. We also appreciate the unceasing support of the University of Bamberg Press.

Finally, we hope that our readers' future research on monitoring, modeling, analysis, simulation, and performance evaluation of next-generation energy networks, energy-efficient systems, and social networks will benefit from the Proceedings of MMB & DFT 2014 International Workshops.

Bamberg, March 2014

Kai Fischbach, Udo R. Krieger, Thorsten Staake
Conference Chairs
MMB & DFT 2014 International Workshops

Invited Talks

Network Analysis Literacy

Katharina A. Zweig

University of Kaiserslautern
Department of Computer Science
Graph Theory & Complex Network Analysis Group
Gottlieb-Daimler-Strasse, Building 48
D-67663 Kaiserslautern, Germany
`http://gtna.informatik.uni-kl.de/en/`

Abstract

Network analysis provides a perspective on how to find and quantify significant structures in the interaction patterns between different types of actors and on how to relate these structures to properties of the actors. It has proven itself to be useful for the analysis of biological and social networks, but also for networks describing complex systems in economy, psychology, geography, and various other fields. Today, network analysis packages in the open-source platform R and other open-source software projects enable scientists from all fields to quickly apply network analytic methods to their data sets. Altogether these applications offer such a wealth of network analytic methods that it can be overwhelming for someone just entering this field. This talk provides some examples demonstrating that not every method can be used for every kind of network or research questions - and that there are even some relations that should not be displayed as networks at all.

Coolhunting for "Honest Signals of Innovation" in Social Media

Peter A. Gloor

Massachusetts Institute of Technology
Center for Collective Intelligence
5 Cambridge Center
Cambridge, MA 02138, USA
pgloor@mit.edu
http://cci.mit.edu/pgloor/index.html

Abstract

This talk describes a series of ongoing projects at the MIT Center for Collective Intelligence with the goal of analyzing the new idea creation process through tracking human interaction patterns on three levels:

On the global level, macro- and microeconomic indicators such as the valuation of companies and consumer indices, or election outcomes, are predicted based on social media analysis on Twitter, Blogs, and Wikipedia. On the organizational level, productivity and creativity of companies and teams is measured through extracting "honest signals" from communication archives such as company e-mail. On the individual level, individual and team creativity is analyzed through face-to-face interaction with sociometric badges and personal e-mail logs.

The analysis is leveraging the concept of swarm creativity, where a small team - the Collaborative Innovation Network (COIN) - empowered by the collaborative technologies of the Internet and social media, turns their creative labor of love into a product that changes the way how we think, work, or spend our day.

The talk introduces the concept of coolhunting, finding new trends by finding the trendsetters, and coolfarming, helping the trendsetters getting their idea over the tipping point. The talk also presents the concept of "Virtual Mirroring" increasing individual and team creativity by analyzing and optimizing five interpersonal interaction variables of honest communication: "strong leadership", "rotating leaders", "balanced contribution", "fast response", and "honest sentiment".

Modelling of the Worldwide Electricity Consumption of ICT

Ward Van Heddeghem, Sofie Lambert, Willem Vereecken, Bart Lannoo,
Didier Colle, Piet Demeester and Mario Pickavet

iMinds-IBCN - Department of Information Technology
Ghent University
Zuiderpoort Office Park, Blok C0
Gaston Crommenlaan 8 Bus 201
B-9050 Gent, Belgium
http://www.ibcn.intec.ugent.be/content/
internet-based-communication-networks-and-services-research-group

Abstract

In this talk, we will provide a detailed description of our modeling methodology for estimating the worldwide electricity consumption of ICT. Our study focuses on three main ICT categories: communication networks, personal computers, and data centers. For these three categories, we assess how ICT electricity consumption in the use phase has evolved from 2007 to 2012. Our estimates show that the yearly growth of all three individual ICT categories (10%, 5%, and 4% respectively) is higher than the growth of worldwide electricity consumption in the same time frame (3%). The relative share of this subset of ICT products and services in the total worldwide electricity consumption has increased from about 3.9% in 2007 to 4.6% in 2012. We find that the absolute electricity consumption of each of the three categories is still roughly equal. This highlights the need for energy-efficiency research across all these domains, rather than focusing on a single one.

Reviewed Papers

SOCNET 2014

Centrality as a Predictor of Lethal Proteins: Performance and Robustness

David Schoch[1,2] and Ulrik Brandes[1,2]

[1] Department of Computer & Information Science, University of Konstanz
[2] Graduate School of Decision Sciences, University of Konstanz

Abstract. The Centrality-Lethality Hypothesis states that proteins with a higher degree centrality are more likely to be lethal, i.e. proteins involved in more interactions are more likely to cause death when knocked off. This proposition gave rise to several new investigations in which stronger associations were obtained for other centrality measures. Most of this previous work focused on the well known protein-interaction network of *Saccharomyces cerevisiae*. In a recent study, however, it was found that degree and betweenness of lethal proteins is significantly above average across 20 different protein-interaction networks. Closeness centrality, on the other hand, did not perform as well.

We replicate this study and show that the reported results are due largely to a misapplication of closeness to disconnected networks. A more suitable variant actually turns out to be a better predictor than betweenness and degree in most of the networks. Worse, we find that despite the different theoretical explanations they offer, the performance ranking of centrality indices varies across networks and depends on the somewhat arbitrary derivation of binary network data from unreliable measurements. Our results suggest that the celebrated hypothesis is not supported by data.

Key words: Network Centrality, Protein Networks, Centrality-Lethality

1 Introduction

With advances in high-throughput analysis, availability of protein interaction data increased dramatically. This provides opportunities to examine interactions and their properties using network analysis. Substantial interest was sparked by Jeong *et al.* [1] who propose that lethal proteins, i.e. proteins causing death if knocked off, tend to have more interactions than non-lethal ones, i.e. they have a higher degree. These findings led to a flurry of follow up studies and a hunt for the centrality best suited to identify lethal proteins [2,3,4,5,6,7,8]. Most of these stuck with the protein-interaction network of *Saccharomyces cerevisiae* used in the original study. Only few studies dealt with different organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans* [9]. In a very recent study, Raman *et al.* [10] reviewed the Centrality-Lethality Hypothesis across protein networks of 20 different organisms. Using a bootstrapping approach, they showed that degree and betweenness centrality of lethal proteins are significantly higher than the network average. In contrast, closeness centrality was found to be less indicative of lethality.

In the following, we reexamine their results, albeit with a variant of closeness centrality correcting for the fact that most of the networks are not connected. Moreover, we use a more detailed evaluation method specifically designed for models with binary outcomes, namely the receiver operating characteristic [15]. Finally, we analyze the robustness of these results when the threshold for high-confidence interactions is varied and discuss theoretical upper bounds for the case when gene attributes are taken into account as well.

2 Methods

2.1 Data

The protein interactions of 20 organisms[3] were obtained from the *STRING Database* (version 9.0). Besides experimentally identified interactions from published literature, the database also contains computationally predicted interactions. Each interaction is given a score which indicates the probability of an actual interaction. We constructed eight networks using $S \in \{600, 650, 700, 750, 800, 850, 900, 950\}$ as lower bounds for the interaction scores for each organism. Lethality data were obtained from the *Database of Essential Genes* (DEG version 5.0).

2.2 Network Analysis

Protein interactions are represented in an undirected graph $G = (V, E)$, where the vertices V represent proteins equipped with a binary attribute indicating lethality

[3] We use the same organisms as in [10], except we choose *D. melanogaster* instead of *S.e.S. typhi*.

and the edges E represent interactions. The cardinalities $|V| =: n$ and $|E| =: m$ denote the number of proteins and interactions respectively. The adjacency matrix $A = (a_{ij})$ encodes the network relation, i.e. $a_{ij} = 1$ if $\{i, j\} \in E$ and $a_{ij} = 0$ otherwise.

For the prediction of lethal proteins we use four standard indices, degree, betweenness, closeness and eigenvector centrality, together with two indices proposed specifically to identify lethal proteins: subgraph centrality [11] and bipartivity [12].

Degree centrality (C_D) is defined as the number of edges incident to a vertex. Betweenness centrality (C_B) quantifies the participation of a node in the shortest paths of the network. It is defined as

$$C_B(v) = \sum_{s \neq t \in V \setminus \{v\}} \frac{\sigma(s,t|v)}{\sigma(s,t)} ,$$

where $\sigma(s,t)$ is the number of shortest paths connecting s and t and $\sigma(s,t|v)$ is the number of shortest paths from s to t passing through v. Closeness centrality (C_C) of a vertex v is defined as the inverse of the sum of its distances to all other vertices in the network,

$$C_C(v) = \frac{1}{\sum\limits_{t \in V \setminus \{v\}} dist(v,t)} .$$

By definition of shortest-path distances, C_C is ill-defined on unconnected networks. Replication confirmed that it was used in [10] nevertheless, which may explain its comparatively poor performance. We therefore use a close variant applicable to both connected and unconnected graphs instead [4],

$$C^*_C(v) = \sum_{t \in V \setminus \{v\}} \frac{1}{dist(v,t)} .$$

Eigenvector centrality (C_E) of a node v is given by the vth entry of the eigenvector corresponding to the largest eigenvalue of A. Again, this formulation is not well-defined for unconnected networks. We therefore calculate C_E for each component separately and scale the values according to the number of nodes in each component. Subgraph centrality (C_S) sums up all closed walks starting and ending at a vertex v. These closed walks are weighted in a way that their contribution decreases as the length increases,

$$C_S(v) = \sum_{k=0}^{\infty} \frac{\left(A^k\right)_{vv}}{k!} = \mathit{trace}(e^A)_v .$$

[4] This variant was proposed, for instance, by Agneessens and Borgatti (presentation at the ASNA 2012 conference)

Bipartivity (β) is defined as the proportion of closed walks of even-length and can be expressed as

$$\beta(v) = \frac{C_{S_{even}}(v)}{C_S(v)} = \frac{\sum_{j=1}^{n} \left[x_j(v)\right]^2 \cosh(\lambda_j)}{C_S(v)} \ ,$$

where $x_j(v)$ is the vth component of the jth eigenvector associated with the eigenvalue λ_j of A. The values of β are confined to the interval $[0.5, 1]$. According to [3], lethal proteins tend to have a low bipartivity score. Therefore we adjust the value by setting it to $1 - \beta(v)$, such that lethal proteins potentially have a higher score.

2.3 Receiver Operating Characteristic

To measure the performance of centrality indices as a predictor for lethal proteins we use the receiver operating characteristic (ROC) [15]. The power of a prediction model can be summarized by the area under the ROC curve (AUC). AUC values are bounded between 0 and 1, where a value of 0.5 is the expected performance of a random classifier and higher (lower) scores indicate a better (worse) prediction than expected by chance.

3 Results

In this section we investigate the predictive power of the six indices to identify lethal proteins. Recall that two of them needed correction to account for disconnectedness. In addition, we examine whether the results are stable with respect to the interaction-confidence threshold S and discuss potential upper bounds for the predictions.

3.1 Prediction Performance

Table 1 shows the AUC values of the six centrality indices for the networks with $S = 700$. In contrast to the results reported in [10], we see that the adjusted closeness performs better than degree and betweenness in most of the networks. However, the efficiency varies strongly across all organisms in general.

Recently it has been argued that the identification of lethal proteins can be improved if centrality indices are combined with further attributes, say from gene expression data [7,13]. In such a scenario, preservation of the neighborhood inclusion preorder remains a necessary condition for a centrality effect to exist. We thus obtain an upper bound on the performance of any classifier by minimizing the non-lethal/lethal inversions over all linear extension of this preorder. Even though the problem of finding minimum inversion extensions is NP-hard [14], we were able to find rankings with an optimal AUC value of 1 for all organisms. This implies that there is a lot of potential for performance improvement when external attributes are incorporated.

Table 1. AUC values for the protein networks with $S = 700$. Bold values indicate the best performance per organism.

Organism	C_D	C_B	C_C^*	C_E	C_S	β
A. bayli	0.80	0.72	**0.84**	0.77	**0.84**	0.82
A. thaliana	0.48	0.54	0.49	**0.56**	0.51	0.47
B. subtilis	0.80	0.70	**0.84**	0.56	0.72	0.72
C. elegans	0.60	0.58	0.61	0.63	**0.65**	0.53
E. coli	0.63	**0.71**	0.68	0.62	0.65	0.53
F. novicida	0.65	0.60	0.68	**0.71**	0.70	0.64
H. influenzae	0.77	0.68	0.78	0.77	**0.80**	0.79
H. pylori	0.56	0.54	**0.58**	0.55	**0.58**	0.57
D. melanogaster	0.63	0.58	0.64	0.60	**0.65**	**0.65**
M. genitalium	0.66	0.61	0.66	0.64	**0.70**	**0.70**
M. pulmonis	0.78	0.73	**0.82**	0.64	**0.82**	**0.82**
M. tuberculosis	0.67	0.63	0.66	0.68	**0.71**	0.70
P. aeruginosa	0.71	0.65	**0.77**	0.74	**0.77**	**0.77**
S. a.NCTC	0.79	0.73	0.85	0.77	**0.86**	0.85
S. a.s.a.N315	0.81	0.73	0.83	0.78	**0.84**	0.83
S. cerevisiae	**0.71**	0.64	0.70	0.70	0.70	0.50
S. pneumoniae	0.72	0.68	0.76	0.71	0.78	**0.79**
S. sanguinis	0.83	0.77	0.87	0.78	0.89	**0.88**
S. typhimurium	0.65	0.62	0.69	0.69	0.70	**0.71**
V. cholerae	0.65	0.61	0.69	**0.70**	**0.70**	0.69

3.2 Robustness

To test the robustness of the results, we varied the confidence threshold S to construct eight networks for each organism. Figure 1 illustrates that prediction accuracy depends heavily on the chosen threshold. Observe that the index producing the highest AUC value varies with S and that the results exhibit high variability in general.

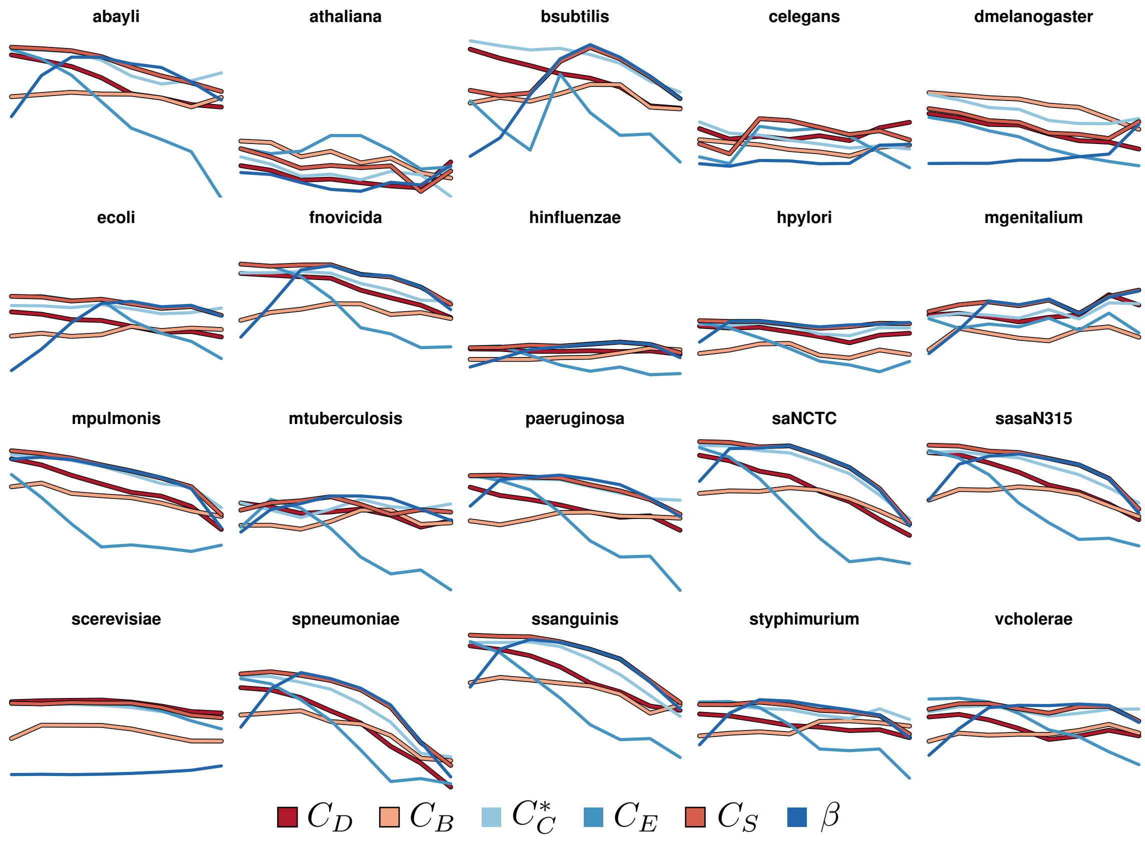


Fig. 1. Performance of the six centrality indices when S is varied from 600 to 950 (shown on the x-axis). The AUC values on the y-axis range from 0.45 to 0.9.

3.3 Discussion

Our reexamination shows that the original results are skewed for two main reasons: inappropriate use of two indices that are ill-defined on disconnected networks, and restriction to a single threshold for interactions. Both are connected to the availability of a finite list of centrality indices, from which instantiations are chosen or new indices are added. Since many of them are defined for connected and unweighted (or otherwise limited classes of) networks, but implementations often output results also for networks outside of this scope, studies need to check carefully whether the aggregate results obtained from such analyses are meaningful.

4 Conclusion

We redesigned and extended a study of Raman *et al.* [10] on the plausibility of the Centrality-Lethality Hypothesis across 20 different organisms. In contrast to [10], we find that (a suitably modified variant of) closeness performs better than degree and betweenness. We also find, however, that the association of centrality and lethality heavily depends on where the line for high-confidence interactions is drawn.

By minimizing the inversions of lethal/non-lethal proteins over linear extensions of the neighborhood-inclusion preorder, we argued that, at least, there is no principled argument *against* a centrality effect. However, in their consideration of purely structural effects, previous results do not provide sufficient support the Centrality-Lethality Hypothesis either.

Acknowledgments. This research was supported in part by *Deutsche Forschungsgemeinschaft* under grant Br 2158/6-1 and the *Graduate School of Decision Sciences.* We are grateful to Stefan Felsner, Franz J. Brandenburg, and Andreas Gemsa for pointing us to the hardness result for the minimum inversion problem.

References

1. Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.
2. Ernesto Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, 2006.
3. Ernesto Estrada. Protein bipartivity and essentiality in the yeast protein-protein interaction network. *Journal of Proteome Research*, 5(9):2177–2184, 2006.
4. Gabriel del Rio, Dirk Koschützki, and Gerardo Coello. How to identify essential genes from molecular networks? *BMC Systems Biology*, 3(1):102, 2009.
5. Xue Zhang, Jin Xu, and Wangxin Xiao. A new method for the discovery of essential proteins. *PLoS one*, 8(3):e58763, 2013.
6. Huan Wang, Min Li, Jianxin Wang, and Yi Pan. A new method for identifying essential proteins based on edge clustering coefficient. In *Bioinformatics Research and Applications*, pages 87–98. Springer-Verlag, 2011.
7. Min Li, Jianxin Wang, and Yi Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Systems Biology*, 6(1):15, 2012.
8. Keunwan Park and Dongsup Kim. Localized network centrality and essentiality in the yeast–protein interaction network. *Proteomics*, 9(22):5143–5154, 2009.
9. Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.
10. Karthik Raman, Nandita Damaraju, and Govind Krishna Joshi. The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Systems and Synthetic Biology*, pages 1–9, 2013.

11. Ernesto Estrada and Juan A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.
12. Ernesto Estrada and Juan A. Rodríguez-Velázquez. Spectral measures of bipartivity in complex networks. *Physical Review E*, 72(4):046105, 2005.
13. Xiwei Tang, Jianxin Wang, and Yi Pan. Identifying essential proteins via integration of protein interaction and gene expression data. In *Bioinformatics and Biomedicine (BIBM) 2012 IEEE International Conference on*, pages 1–4, 2012.
14. Eugene L. Lawler. Sequencing jobs to minimize total weighted completion time subject to precedence constraints. *Annals of Discrete Mathematics*, pages 75–90, 1978.
15. Charles E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, pages 283–298, 1978.

Proposal for Heuristics-based Refinement in Clustering Problems

Antonio A. Gentile[1], Angelo Corallo[1], Cristian Bisconti[1] and Laura Fortunato[1]

Dept. of Innovation Engineering, University of Salento, 73100 Lecce, Italy,
antonio.gentile@unisalento.it,
http://emi.unisalento.it/sna

Abstract. Community detection in networks has recently obtained a huge interest in both natural and social sciences, for its variety of implications and applications. Several algorithms and strategies have been proposed up to now, mainly focusing on speed optimization or on the quality of the final clustering obtained. The main scope of this paper is to bridge these two approaches, via the introduction of heuristic schemes, which may be intended as a ranking of nodes inside a graph, indicating those who pose major problems in community assignment. By performing the slowest, most effective algorithms (refinement) on only a small fraction of the whole network, this approach is made applicable to huge networks.

Key words: clustering, community detection, heuristic, graph partitioning

1 Introduction

Identifying dense structures inside a network may be of crucial importance for a wide variety of reasons. In fact, these clusters[1] may correspond to functional/logical units, or social communities of the network [1]. Even an evaluation of how strong it is, in a network, the tendency to form communities (without actually identifying them) may be of practical interest itself, as this patterning is often connected to robustness and stability [2].

[1] In the literature, for this same concept, also the following terms are equivalently used: *communities, groups, modules, partitions.*

This variety of interesting applications has led to the intense development of algorithms, aiming to solve automatically the problem of finding clusters inside a network, or evaluating if a good partitioning is indeed possible [3]. The purpose of this paper is to provide a strategy, enabling to bridge approaches based on time consuming algorithms, and faster methods: the first ones being devoted to small networks where they can produce reliable results in a limited amount of time (thus making superfluous to adopt faster methods), whereas fast algorithms are the only feasible choice for large networks. We envisage that it is possible to overcome this difficulty, providing a general description of a strategy, suitable for this purpose. In particular, that there may be no need of running a 'refinement step' based on a slow method on the whole network, by focusing the attention on those nodes only, which are identified as *critical* in an intermediate step.
In Sect.2, after a brief introduction on the framework of our proposal, we will provide a detailed assessment of general features and applicability of our multi-step scheme, and discuss how to detect critical nodes. Characteristics and a first testing of the heuristic proposed, with computational results obtained from real-world networks, will be illustrated in Sect.3. Some remarks and outlines of future developments conclude this work.

2 Framework and Methods

In this work, we are going to use concepts and metrics derived from graph theory (see [4]), assuming the following:

Proposition 1. *The network to analyze can be represented by a (un)directed, (weighted) graph $\mathcal{G}$.*

Where:

Definition 1. *An* ***undirected (unweighted) graph*** *is an ordered pair $\mathcal{G}(V, E)$ of the finite set of n elements V (each v_i called a 'vertex' of $\mathcal{G}$) and the associated 2-sets E, composed by m unordered subsets of two elements*[2] *of V (each subset $\{v_i, v_j\}$ being called an 'edge' of $\mathcal{G}$).*

The adoption of a *simple graph* definition leaves apart the more complex cases of clustering in networks represented by *digraphs* and *multimodal graphs*. Specifically, the focus will be on weighted undirected graphs:

[2] Notice that we are allowing *loops*

Definition 2. *An undirected weighted graph is a graph $\mathcal{G}$ as in Def.1, including additionally a **weight function** w defined on its edge set, i.e.* $w : E \to \mathbb{R}$

Among the various possible representations of $\mathcal{G}$, we will mainly refer in the following to:

Definition 3. *The $n \times n$ matrix A with entries $A_{ij} = 0$, if $\{v_i, v_j\} \notin E$, or $A_{ij} = w(\{v_i, v_j\})$, if $\{v_i, v_j\} \in E$ is the **adjacency matrix** associated with the graph $\mathcal{G}$.*

As pointed out in the introduction, the scope of this work is proposing a method, to reduce the computation time required by the straightforward application of clustering algorithms to a network, without a consistent degradation of the results. Therefore, our efforts are in the framework of the vast literature dealing with the automatic clustering of a network.
Among the various strategies developed, following [5] we outline two main directions: *graph partitioning* and *community detection* algorithms. The main difference among them is that, in the first case, number and size of the clusters to be retrieved are defined a priori in the problem: whatever the actual structure of the network, the scope is to find the optimal partitioning according to given parameters. Community detection is understood, instead, as more 'network-driven': in general, no strict input (such as: number or cardinality of final clusters), is given on the final partitioning solution, which also addresses the issue of the *quality* achievable by the clustering procedure. Considering that both these directions can be pursued with our scheme, according to the specific implementation chosen case-by-case, we will adopt the general:

Definition 4. *By 'clustering' or 'community detection' in graph $\mathcal{G}$, we hereafter mean an optimal partition problem*[3] *of finding $\mathcal{C}_k$ components (i.e. subgraphs) of $\mathcal{G}$, where their number k may be an input of the problem, or left as a free parameter.*

The optimal components of the partition P to be retrieved must satisfy the properties:

1. $1 < k < n$, which excludes limit cases [4];
2. $\bigcup_k \mathcal{C}_k = \mathcal{G}$;

[3] We stress not to confuse it with the *graph partitioning*, term commonly used for a specific clustering problem, see [3]

[4] I.e. $\forall k : \mathcal{C}_k \neq \oslash$, and the particular cases $k = n$ and $k = 1$ would be a trivial partitioning

3. $\forall k \neq l : \mathcal{C}_k \cap \mathcal{C}_l = \oslash$ (no overlapping communities);
4. the partition defined optimizes a 'quality function' [5].

About the quality function to be optimized (Pt.4), this may be defined in various ways and for our procedure, it is of no fundamental importance which in particular is chosen. The most popular example of such a function is the *modularity* [6], which will be used several times in the following. In fact, its maximization is considered to be reliably connected with the optimization of the partition quality [7]. The interested reader will find out a survey of other possible such functions in [3]. It may be observed that Pt.4 is rather strict, since in some procedures it would be difficult to define such a cost/quality optimization (e.g. in *hierarchical clustering* several solutions for the partitioning may be chosen as optimal, according to different considerations): in the next paragraph we will explain why it is reasonable to assume it in a multi-step scheme.

A classification of algorithms based on their specific scope has already been given. On a different basis, one could distinguish among general classes, grouped according to the algorithm performance and method. A first one, devoted to capturing the global picture of the network clustering, aiming at a fast solution of the clustering problem given, which especially suits large networks. Such algorithms will be generically indicated in the following as *coarse grain*, since in general they use global metrics as the figure of merit to optimize, and often embed approximated methods, thus potentially leading to a relatively high rate of misclassified nodes (e.g. see [7]). A first set of good examples for this case is given by *optimization methods*[6]; e.g. those involving *E/I ratios* [8], information-compression [2] ..., Hamiltonian-like quantities (spin-hamiltonians, *modularity*, ...). For each of these measures, several different algorithms do exist, according to the optimization schedule which is followed. Taking again modularity as a paradigm, both greedy [9] and local optimization [10] have been investigated, as well as other techniques not reported here [11]. Spin energetic schemes have also proven useful in multiresolution approaches [12]. Another vast set of available algorithms is known as *block modeling* [3]: statistical methods again aiming at the optimization of a quality function defined in terms of an *image graph*, which can recover some of the optimization methods seen above, for particular choices in the model parameter.

[5] Which may equivalently be a 'cost function': for brevity, the same expression will here refer to both cases

[6] Following the classification given in [1]

On the opposite side, fine grain algorithms, in particular those involving metrics at the node/edge level[7], or *hierarchical* structures: in this case, the aim is a precise assignment of the single nodes to the various communities. Moreover, these *refinement* algorithms frequently adopt 'exact' methods, for the optimization task they deploy.

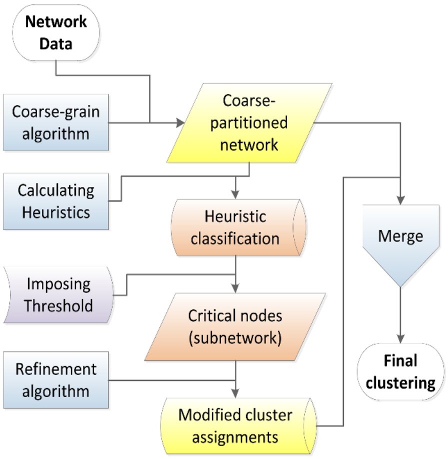


Fig. 1. Flowchart of the multi-step method proposed in the text. In blue/violet are distinguished the operational steps, where, in particular, violet ones involve the tuning of free parameters, which are specific of this proposal. Other colors are referred to data (in red is emphasized the core of our proposal): database symbols are used to indicate data, which do not need to be structured in form of a graph.

2.1 The multi-step scheme

The scope of our approach is to bridge the two different classes of coarse-grain and refinement methods. At the moment, in fact, it is unusual to find a refining step adopting a different metric: the norm is the straightforward application of a single step algorithm [3], or different optimization schedules for the same quality function (see [14] for an example with *modularity*). There is a logical reason behind this tendency: refinement algorithms are unfeasible for large networks, and are already the best strategy available for smaller cases. Our contribution is the proposal of *heuristic metrics*, having both low computational time-complexity, and a good efficiency in classifying the nodes according to their degree of membership to possible communities. We call it 'heuristic' because, as discussed in the

[7] Like the *edge betweenness*, *information centrality*, other cost functions, directly referred to the network structure (like for the Kernighan-Lin approach [13]), or to real-world analogies [1]: *current-flow*, *message-passing*, ...

following, the metric chosen not only draws on the characteristics of the network analyzed, but must rely on some 'preliminary' clustering results, as computed via coarse algorithms.

These metrics enable the adoption of a scheme including the following elements:

1. a *coarse grain algorithm* for the initial clustering guess,
2. an efficient, *heuristic metric* for the retrieval of a reduced set of nodes, requiring further analysis upon cluster assignment,
3. a *refinement algorithm*, to be run on the nodes produced by the previous step, i.e. a fraction of the initial graph, to improve the 'quality' of the final partitions.

These three main elements, along with some other features which will be introduced in the text, are in Fig.1, which shows the global multi-step structure.

It is now more evident, how assumptions in Prop.1 and Pts.1-3 of Def.4 are required, for the following discussion to make sense. Even if, in the following, we will assess to what extent Pt.3 may be (partially) relaxed, though leading to interesting applications. Pt. 4, instead, eases the adoption of a multi-step scheme and is no fundamental assumption. Indeed, as further outlined below, for this whole procedure to be applicable, there is merely the need for a (single) 'starting-point' partition, as produced by the coarse-grain algorithm, and a refinement algorithm with a few specific features. These algorithms may not include an explicit cost optimization process: in turn, if present, this optimization eases a quantitative comparison among different solutions, and in some cases is necessary to avoid assumptions a priori on the structure of the clustering (e.g. number and sizes of the clusters, see [15]), thus greatly losing in the generality of the method. Therefore we assumed a cost function in Def.4.

The very same introduction of a refinement concept brings along the problem of identifying a measure, able to evaluate the 'quality' of a clustering (i.e. an *absolute* measure), or at least to compare different clustering solutions (i.e. a *relative* measure). To the authors' knowledge, there is no widely accepted definition of a distance between different clustering solutions of the same graph, say P and P', even if a few proposals [3] arose for both the absolute (e.g. the *modularity*) as well as the relative case (e.g. the *performance*). A general discussion about the problem is clearly outside our scope, additional information can be found in [16],[17],[18].

As the first, a simplicistic assumption: the best clustering solution, which the fine grain algorithm is capable to find when applied to the whole network, *is* the best possible approximation to the ‘true’ clustering of the network (if any is known). Our procedure aims to approximating the clustering, which would be provided if the fine-grain analysis chosen was to be performed on the whole network: this will be implicitly the ultimate target, without questioning further the effectiveness of the fine-grain step.

As the second, for the distance among partitions provided at different steps in our procedure, we adopt the simple and intuitive concept in [19]: *“the minimum number of elements that must be deleted from (a graph)... so that the two induced partitions... restricted to the remaining elements are identical”*. Calling this number $n_D(P, P')$, a slight modification produces the definition of the *partition-distance* $D(P, P')$ as the ratio between nodes classified differently, and the size of the graph:

$$D(P, P') := n_D(P, P')/n \tag{1}$$

The lower the distance between a certain partition, and the one provided for the whole network by the refinement algorithm chosen, the better the partition identified. Therefore, D shall be used in numerical experiments testing the utility of our proposal.

Proposition 2. *Our multi-step procedure requires a few qualitative hypotheses to be made, for the strategy to be effective:*

1. *the refinement algorithm chosen must perform displacements of single nodes;*
2. *the heuristic metric chosen must perform as a good figure of merit, in quantifying the ‘criticality’ of the nodes in the network;*
3. *however chosen, the fastest method (eventually approximate) to compute the metric as above should outperform, in time-complexity, the refinement clustering algorithm.*

Further discussing these hypotheses, the first is obviously derived from the necessity, once a node-ranking has been established, to analyze and eventually modify the cluster attribution of specific nodes. If this is intrinsically impossible, given the way the refinement algorithm works, the whole scheme of finding critical nodes to refine may prove useless.

The second statement emphasizes how a key point, in our scheme, is to choose a heuristic metric, able to reliably identify nodes which are likely to be misassigned (i.e. the ‘critical nodes’). A perfectly efficient metric should rank first *only* nodes which will be misassigned by the coarse grain algorithm. Clearly, given that a variety of algorithms could be used as coarse-grain, this ‘perfect efficiency’ is indeed a relative concept, and independently from the refinement algorithm there is no way to define it. Moreover, an important feature, this heuristic should exhibit, is a reduced computational complexity, as stated in the third place. This is evident from a simple analysis of the procedure structure: for an alternative clustering procedure to be competitive, with respect to the refinement algorithm, all of its steps must be (much) faster to compute. Yet, if the fast algorithm employs itself heuristics, adopting the same metric for the second step of the procedure would add no further information to the problem: this consideration must as well be taken into account.

A first naive approach, for retrieving the critical nodes of the network, could be to adopt typical node metrics employed in network theory: *centrality* measures [20], e.g. the *degree*, *closeness*, *betweenness* ... of the nodes. The point, in this choice, would be the adoption of well-known metrics, for which a wide variety of efficient computing algorithms has been proposed [21]. However, there are two major drawbacks. The first is the implicit assumption, that the refinement should involve the most important nodes. Therefore, the strategy subtended to this approach is: "ensure via the refinement that most important nodes are correctly classified", rather than "efficiently retrieve nodes which are likely to be misclassified". As a second problem, notice how some of these metrics are ruled out by Pt.3 in Prop.2; e.g. the fastest algorithms known [22] for computing the betweenness of an unweighted graph are $\mathcal{O}(nm)$, which is acceptable, whilst for the exact computation of the closeness are required algorithms $\mathcal{O}(nm+n^2log\ n)$, a complexity worse than those of some fine-grain clustering algorithms [23].

Given these preliminary considerations, we will not ponder further the possibility to use this kind of metrics, dedicating instead the rest of this paragraph to the development of a heuristic suitable for our purposes, i.e. satisfying the requirements in Prop.2.

Let us introduce a few qualitative statements. As the first, the assignment of a node to a cluster is determined by how its links (i.e. edges) to neighbour nodes are distributed [3]. Thus, the (weighted) edges from/to the node to be classified must play an important role in our heuristics: this leads to introducing the node degrees. In order not to relate the heuristic to the importance of the node, some normalization factor must be introduced. Leaving further considerations about the (eventual) *direction* of the edges, we will use a total 'symmetrized' degree of

node j, $d_T(j) := \sum_i (a_{ij} + a_{ji})/2$.
Because of the specific aim of the heuristic, and given the hypothesis of computing it only after a first coarse assignment of nodes to clusters has been made, one is able to distinguish among edges *inside* or *outside* a given cluster. To improve the readability of the formulas below, we here introduce the binary function com with values in $\{-1, 1\}$, defined for each couple of indexes (i, j) denoting nodes, which belong to the graph:

$$com(i,j) := \begin{cases} -1 & \text{(if } i \text{ and } j \text{ belong to different communities)} \\ +1 & \text{(if } i = j \vee \text{ if } i \text{ and } j \text{ belong to same cluster)} \end{cases} \tag{2}$$

Another important figure of merit for our heuristics is:

$$Q = \frac{\delta(\mathcal{G})}{\Delta(\mathcal{G})} \tag{3}$$

where $\delta(\mathcal{G})$ and $\Delta(\mathcal{G})$ are respectively the minimum and maximum degree of the nodes in graph $\mathcal{G}$. Q will play the role of a normalization factor in the following.

All the elements for formulating a proposal, for the metrics satisfying the prerequisites as above, have now been discussed. We claim that a 1^{st} order heuristic metric, suitable for quantifying the criticality of node j, is of the form:

$$H_1(j) = \frac{1}{2d_T(j)} \sum_i (a_{ij} + a_{ji})\, com(i,j) \tag{4}$$

while for the 2^{nd} order heuristic we suggest:

$$H_2(j) = \frac{1}{2d_T^2(j)} \sum_{i \neq j} (a_{ij} + a_{ji}) \;\; com(i,j) Q \;\, d_T(i) H_1(i) \tag{5}$$

A formal remark: distinguishing them as 1^{st} and 2^{nd} order, there is no assumption about the order of magnitude of the expected ‘error’ (here intended as the percentage of misassigned nodes, having a heuristic metric higher then the threshold). In fact, these expressions refer to the width of the network sample taken into account for each node: the edges shared *with* its neighbour nodes in the 1^{st} case, and also all edges shared *by* its neighbour nodes in the 2^{nd}. Before moving on to a discussion about the effectiveness of the metrics chosen, a few preliminary comments.

The first order heuristic is bounded, as $-1 \leq H_1 \leq +1$, thus it may be interpreted as a normalized measure of the *correlation* of the node with its cluster of assignment, disregarding any feature of its neighbor nodes. Evidently, a positive correlation is here an index of robust assignment (high ratio of links inside-cluster, or 'internal', versus links outside), whereas negative correlations indicate misassignment. It is worth further commenting the limit cases: if H_1 is equal to $+1$ (-1), this would mean that all of the neighbour nodes of j belong to a cluster, same as (other than) the cluster of j. Therefore, even if the possibility $H_1 = -1$ is retained theoretically, in practice it should be implausible to find nodes, where the coarse-grain algorithm did perform so bad.

The second order heuristic is conceived as a 'normalized' metrics as well: it has the same lower and upper bound as H_1. Comparing (5) with (4), it is easy to show why. The summation in (5) is performed over $n_j - 1$ elements, where n_j is the cardinality of the set $\mathcal{N}_j$ of unique nodes satisfying the condition $com(i, j) = 1$, with j fixed. Each of them includes two factors:

$$H_1(i) \tag{6}$$

$$M := Qd_T(i)/d_T(j), \tag{7}$$

which are both less than one in module (see in (3) how Q is defined). Replacing both of them with this limit case, the discussion can be reduced to the same as for $H_1(j)$, which completes the proof: $-1 \leq H_2 \leq +1$. Qualitatively, re-introducing in (5) the heuristic H_1 accounts for the cluster assignment of neighbour nodes: the stronger the connection of a neighbour node i to its own cluster, the higher we expect its contribution to the (mis)assignment score of analyzed node j, if $com(i, j) = +1$ (-1). The factor in (7), instead, can be interpreted as a measure relating the contribution from node i to its relative 'importance' in the network, compared to node j (thus the presence of Q). When using a combination of the two heuristics proposed, this same term reduces the contribution from H_2, compared to H_1: whenever $d_T(i)/d_T(j) << 1$, the quantity in (7) will thus not exceed the square of this ratio. This is yet another good reason to name H_2 a '2^{nd} order heuristic'.

If self-loops are present in the graph, they contribute to H_1 (as they increase the number of internal links), but they play no role in H_2: as we have seen, this last term was intended to emphasize the role played by neighbour nodes on the assignment, and therefore the case $i = j$ has been explicitly excluded by the summation in (5).

Another interesting point to analyze is how to use the two heuristics introduced. A first possibility is of course to refer singularly to each of the heuristics: Eq.(4), so to emphasize the role played by the importance of the node investigated itself, or Eq.(5) in order to focus on the *nearest-neighbourhood*. However, we found that the most profitable strategy is to combine the two heuristics in one single figure of merit. The most natural way, to perform the combination, is to sum the 1^{st} and 2^{nd} order contributions:

$$H(j) := \alpha H_1(j) + (2-\alpha) H_2(j) \tag{8}$$

with $\alpha \in [0,2]$. In the following, we will always refer to the simplest case with $\alpha = 1$. They certainly may be considered more complicated combinations[8]. However, no substantial difference in the results has been observed for the simplest cases, therefore, in the following of this proposal, we kept using (8) with $\alpha = 1$. Before moving on to further discuss applications and features of the method introduced, it is worth a brief comment about the mutual relationships between Formulas (8), (5) and the simplest metric (4). The introduction of heuristics as above may be regarded as a 'mean field like' procedure, where only pairwise, nearest neighbour interactions are considered (which is the case, for example, in Ising models), for deriving a quantity H which can be interpreted as a *potential*, once changed in sign. One may recall other *Hamiltonian*[9] approaches to clustering problems: e.g. see [24] for the application of a 'Potts model' (from which modularity-based methods themselves derive), and [25] for an 'Ising model'. Indeed, with a terminology drawing on this parallel, a key difference in our approach is that we are defining and using *local* potentials, whereas the traditional approach involves the definition and optimization of a *global* potential.

2.2 Discussion about implementation

We are now left with checking the respondency of our proposal for a heuristic, to the requirements stated in Pts. 2-3 of Prop.2. Taking in consideration the general formulas (4) and (5), for the case of an undirected graph, it is easy to see that H_1

[8] As an example, we have tested also a weighted formula, involving the degrees of the nodes and preserving the normalization also in the compound heuristic: $\alpha = \frac{d_T(j)}{d_T(\mathcal{N}_j)}$.

[9] Where the Hamiltonians are defined starting from analogies with physical models, and identifying explicitly some of the terms, as dependent on the community assignments [3].

has a complexity of $\mathcal{O}(m)$, if starting from a graph in the form of ordered *edgelist*[10]. The two heuristics have redundant terms, e.g. H_1 for the nodes can be stored in an additional 'heuristic vector', and read-out for computing H_2. If this is the case, H_2 can be computed in additional $\mathcal{O}(n+m)$ time. This is a reasonably good result: e.g. one of the fastest coarse algorithms for solving clustering problems runs with complexity $\mathcal{O}(n+m)$ on sparse graphs (this, along with further examples, can be found in [3]). Additionally, operations leading to the heuristics' complexity are very basic, thus we envisage very low factors.

In order to perform a test for the multi-step scheme, following also Fig.1, two elements are required to be explicitated:

1. A coarse grain algorithm for the first step. We chose to use the *fast Newman* (FN) approach [6], because it is of widespread adoption in the literature[11] and in several network analysis softwares. This approach is based on *modularity* as a quality function, with a greedy modularity optimization, as suggested in [9]. Within this implementation, the algorithm is known to run in $\mathcal{O}(n\ log^2 n)$ on sparse graphs.
2. A refinement algorithm for the final step. In this case we used a modified *Girvan-Newman* (GN) method, based on the *edge betweenness*: a perfect example of an algorithm unfeasible to be used straightforward for large networks, as it requires $\mathcal{O}(n^3)$ time (sparse case). The original version of this algorithm was not intended to perform single node re-assignments [27], so that it is here slightly modified, even if keeping the working principle. In particular, where the original paper had the step '*calculate betweenness for all edges...*', here the calculation is performed only on edges linking couples of nodes, of whose at least one is *critical*. Additionally, the last edge to be removed, before a node is isolated, is also the one ruling the community assignment.

Notice that, in this way, the refinement chosen is allowed to eventually shrink the number of clusters composing the final partitioning, but explicitly avoids the creation of new clusters. Such a possibility exists, but we will not discuss it in detail here: in the spirit of general applicability of the scheme proposed, indeed, it is superfluous to investigate the capabilities of a particular implementation, that

[10] If not, an additional step with complexity $\mathcal{O}(m\ log\ m)$ must be taken into account

[11] Its combined simplicity and robustness make the FN method still very popular, even if several works have started to point out its ineffectiveness for specific cases [7], [26].

are not available in other refinement procedures (e.g. if a *Kernighan-Lin* method was to be chosen [13]).

3 Preliminary tests

This paragraph is devoted to show how the particular implementation of a multi-step scheme (as outlined in Sect.2.2) works for a real case, and in particular to test if the heuristics, introduced so far, are capable of satisfying the requirements stated at the beginning of this section.

A test-case. We have chosen to use the split of a *karate club* in two different 'communities', studied in [28] and often used as a benchmark in the literature about community detection. This example fits well a preliminary, qualitative discussion, because it is small enough (n=34) to let us follow in detail the performance of the heuristics, yet it is complex enough to pose difficulties for the fast coarse algorithm chosen [5]. In Fig.2 we plot the values of the heuristics H_1 and H_2 for all the vertices of the karate club network, after a coarse assignment of clusters has been performed through the application of the FN algorithm. It is immediately evident how, as envisaged in Sect.2, in most cases the values of H_1 are much bigger (in module) than the corresponding H_2 values, thus correctly configuring this 2^{nd} order metric as a minor correction in our final compound heuristic. Another interesting feature is that almost all of the nodes have positive values of both $H_{1,2}$: this confirms how the coarse algorithm chosen performs good enough in this test. Finally, we can also state that our core claim is satisfied: the node #10, known to be misclassified by the coarse algorithm [6], is the one scoring worse, and even has a negative H, as shown in Fig.2. Therefore, this first test encourages the adoption of our heuristic to scout critical nodes, which are likely to be misassigned. Within this test, the refinement algorithm has been run on all, and only, those nodes having negative values of H. That is on the vertex #10 alone. The GN refinement correctly classifies this node, displacing it into the 'right' cluster. Recalling (1), the coarse method has in this case a distance of $D \cong 0.029$ from the partition found by our scheme: this distance can be understood as the *improvement* provided for the solution.

Before moving on to the conclusions of this work, there is one more remark to be done. The idea of measuring the 'strength of membership' of the vertices to their clusters is not unprecedented, in the realm of graph partitioning and community

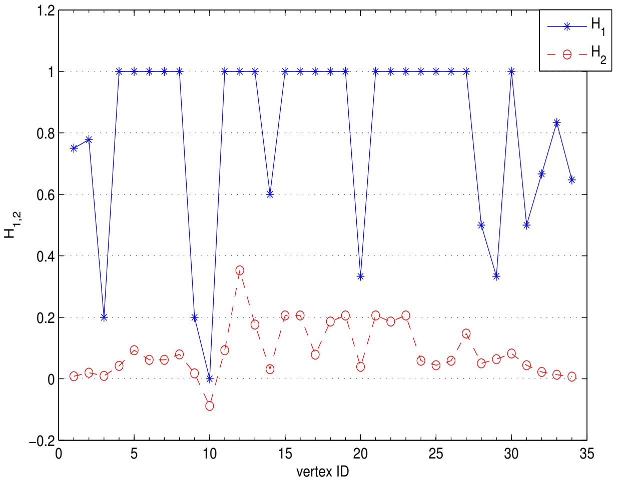


Fig. 2. Calculation of the 1^{st} and 2^{nd} order heuristics H_1 and H_2, for the case of Zachary's karate club, outlined in the text. A legend is provided for distinguishing the two cases. Vertices IDs are as in [27]. Node #10 is considered critical by analyzing results of the FN coarse algorithm cited in the text.

detection. For the case of non-overlapping communities, in fact, it is possible to find in the literature an older attempt to quantify this strength, via the elements of the *modularity matrix* eigenvector [5]. This approach led to results which are indeed complementary to the ones presented here. In fact, considering the membership values for the karate club case (analyzed in the cited work, where they are linked to the modularity eigenvector linked to the modularity eigenvector), notice how the measure in [5] is suitable for identifying a few vertices, which are strongly connected to their cluster, whereas our heuristics exploit those few vertices, with the weakest attribution to a cluster in particular.

4 Conclusions

Summarizing the main results of this work: we have proposed the adoption of a multi-step scheme, to improve the results of clustering algorithms, with a particular focus on community detection. This scheme basically includes: the adoption of a (state-of-art) fast, coarse algorithm for the first step; an accurate (state-of-art) refinement algorithm, eventually adapted for this specific purpose; to bridge these two elements, a novel set of heuristic metrics. These last ones are the core of the proposal: they are intended to scout those nodes potentially tricky in the cluster assignment, and thus worth to be analyzed by the refinement step. We have shown, also with the aid of a test-case, that the heuristic introduced satisfies the requirements of being computable with low time-complexity, and may efficiently retrieve those nodes which turn out to 'deceive' the less accurate algorithms.

References

1. M. Newman, "Communities, modules and large-scale structure in networks," *Nature Physics*, vol. 8, no. 1, pp. 25–31, 2011.
2. M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.
3. S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.
4. D. B. West *et al.*, *Introduction to graph theory*, vol. 2. Prentice hall Englewood Cliffs, 2001.
5. M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.
6. M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, p. 066133, 2004.
7. B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, no. 4, p. 046106, 2010.
8. D. Krackhardt and R. N. Stern, "Informal networks and organizational crises: An experimental simulation," *Social psychology quarterly*, pp. 123–140, 1988.
9. A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.
10. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
11. M. Newman, *Networks: an introduction*. Oxford University Press, 2009.
12. P. Ronhovde and Z. Nussinov, "Multiresolution community detection for megascale networks by information-based replica correlations," *Physical Review E*, vol. 80, no. 1, p. 016109, 2009.
13. B. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell system technical journal*, 1970.
14. M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.
15. A. Pothen, "Graph partitioning algorithms with applications to scientific computing," in *Parallel Numerical Algorithms*, pp. 323–368, Springer, 1997.
16. R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *Journal of the ACM (JACM)*, vol. 51, no. 3, pp. 497–515, 2004.
17. H.-P. Kriegel and M. Pfeifle, "Measuring the quality of approximated clusterings," in *BTW*, vol. 5, pp. 415–424, 2005.

18. C. Robardet, F. Feschet, and N. Nicoloyannis, “An experimental study of partition quality indices in clustering,” in *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD ’00, (London, UK), pp. 599–604, Springer-Verlag, 2000.
19. D. Gusfield, “Partition-distance: A problem and class of perfect graphs arising in clustering,” *Information Processing Letters*, vol. 82, no. 3, pp. 159–164, 2002.
20. S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.
21. U. Brandes and T. Erlebach, *Network analysis: methodological foundations*, vol. 3418. Springer, 2005.
22. U. Brandes, “A faster algorithm for betweenness centrality*,” *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.
23. K. Okamoto, W. Chen, and X.-Y. Li, “Ranking of closeness centrality for large-scale social networks,” in *Frontiers in Algorithmics*, pp. 186–195, Springer, 2008.
24. J. Reichardt and S. Bornholdt, “Detecting fuzzy community structures in complex networks with a potts model,” *Physical Review Letters*, vol. 93, no. 21, p. 218701, 2004.
25. S. Liu, L. Ying, and S. Shakkottai, “Influence maximization in social networks: An ising-model-based approach,” in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 570–576, IEEE, 2010.
26. A. Kehagias, “Bad communities with high modularity,” *preprint arXiv:1209.2678*, 2012.
27. M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.
28. W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of anthropological research*, pp. 452–473, 1977.

Information Dissemination Processes in Directed Social Networks

Konstantin Avrachenkov[1], Koen De Turck[2],
Dieter Fiems[2] and Balakrishna J. Prabhu[3]

[1] Inria Sophia Antipolis, France
[2] University of Gent, Belgium
[3] LAAS-CNRS, France

Abstract. Social networks can have asymmetric relationships. In the online social network Twitter, a follower receives tweets from a followed person but the followed person is not obliged to subscribe to the channel of the follower. Thus, it is natural to consider the dissemination of information in directed networks. In this work we use the mean-field approach to derive differential equations that describe the dissemination of information in a social network with asymmetric relationships. In particular, our model reflects the impact of the degree distribution on the information propagation process. We further show that for an important subclass of our model, the differential equations can be solved analytically.

Key words: Directed graphs, mean-field, configuration model

1 Introduction

We develop mathematical models for the dissemination of information on directed graphs and investigate the influence of parameters such as the degree distribution on the dynamics of the dissemination process. The directed graph model, as opposed to the undirected model, is better suited for networks like the one of Twitter because of the asymmetric relationship that exists between different users. Specifically, in the Twitter network, a user can choose to receive the tweets—in other words, become a follower—of one or more other users by subscribing to their accounts. Certain users, celebrities for example, have several millions of followers who follow their tweets. These users do not necessarily follow the tweets of all of their followers which results in an asymmetric relationship between users. This asymmetry is modelled by a directed graph in which an outgoing edge is drawn from a user to each of its followers. An edge in the opposite direction from the follower to the user need not always exist and is drawn only if the user subscribes to the channel of this follower.

A hashtag is a word or a phrase prefixed by # and used in social networks as a keyword. The prefix facilitates the search for conversations related to the prefixed word or phrase. The typical life cycle of a hashtag closely resembles an epidemic. In the first phase the interest in the hashtag grows as users generate tweets containing this hashtag. These tweets are received by followers who then either retweet them or generate new tweets with this hashtag. The number of users tweeting this hashtag ("infected" users) grows as a function of time during this phase. At a certain point in time, the interest reaches its zenith and starts to wane as users move on and get interested in other events. The second phase begins at this point in time as users stop tweeting this hashtag (or, "recover"), and the number of infected users decreases.

In this work we use the mean-field approach to derive the differential equations which describe the lifecycle of hashtags or, in other words, the process of information dissemination. We obtain a couple of differential equations which describe the evolution of the fractions of infected and recovered persons. As a model for the underlying network we take the Configuration-type model for directed graphs [1]. While epidemics have been widely studied on undirected graphs, there is a hardly any analysis of the epidemic-type processes on directed networks. In [2,3], the mean-field approach has been applied to the analysis of epidemics on an undirected configuration-type graph model. In [4], the effect of network topology has been analysed in the case of undirected graphs. In particular, the authors of [4] applied their general results to analyse the Erdös-Rényi and preferential attachment random graph models. An interesting approach combining a decomposition approach with two-state primitive Markov chain has been proposed in [5] for undirected networks with general topology. For an overview of various results about epidemic processes on undirected networks we refer the interested reader to the books [6,7,8].

2 The mean-field model

Consider a network of N nodes structured according to the Configuration-type model for directed graphs [1]. The in-degree and out-degree of the nodes are drawn from a distribution $f(k,l) = \mathbb{P}(K = k, L = l)$, defined on the bounded set $\mathcal{D} = \{(k,l) : 0 \leq k \leq \hat{K}, 0 \leq l \leq \hat{L}, (k,l) \neq (0,0)\}$, where the first (resp. second) index corresponds to the in-degree (resp. out-degree) and $\hat{K}$ (resp. $\hat{L}$) is the maximal in-degree (resp. out-degree). In the remainder, we always assume that $\mathbb{E}K = \mathbb{E}L$; in a network every outgoing link is an incoming link of some other node. A generic node with in-degree k and out-degree l shall be referred to as a (k,l)-node.

In order to generate a Configuration-type graph, each node first draws its in-degree and out-degree (resp. called incoming stubs and outgoing stubs) from the given distribution. In order to ensure that the total out-degree is equal to the total in-degree, stubs can either be added or removed. Each outgoing stub is then connected to a free incoming stub picked uniformly at random. After connecting all the stubs, the self-loops and multiple edges between nodes are removed. It was shown in [1] that a graph generated using this procedure leads to a Configuration-type graph.

Each node in the network can be in one of the three states : infected, recovered, or susceptible. An infected node infects its susceptible neighbours after an exponentially distributed time with intensity λ. Note that, the infection is spread simultaneously along all the outgoing edges and not just one edge at a time. The simultaneous dissemination along all outgoing edges models the spread of tweets on Twitter. An infected node recovers after an exponentially distributed time of rate ν, at which time it stops spreading information in the network.

We shall be mainly interested in a large-population model, that is when $N \to \infty$. This assumption simplifies considerably the analysis of the dissemination process while being realistic[4].

Let $i_{k,l}(t)$ (resp. $r_{k,l}(t)$) denote the fraction of infected (resp. recovered) (k,l) nodes at time t. The following result describes the dynamics of these two quantities.

Theorem 1. *Let* $i_{k,l}(0) > 0$ *for some* $(k,l) \in \mathcal{D}$. *Then,* $\forall (k,l) \in \mathcal{D}$,

$$\frac{di_{k,l}(t)}{dt} = \lambda k(f(k,l) - i_{k,l}(t) - r_{k,l}(t)) \frac{\sum_{k',l'} l' i_{k',l'}(t)}{\sum_{k',l'} l' f(k',l')} - i_{k,l}(t)\nu, \tag{1}$$

[4] Twitter has approximately 200 million registered users (source: Wikipedia).

and
$$\frac{dr_{k,l}(t)}{dt} = i_{k,l}(t)\nu. \tag{2}$$

Proof (Sketch of proof). Let $N^{(N)}_{k,l}$ be the total number of (k,l) nodes in a network of N nodes, and let $I^{(N)}_{k,l}(t)$ (resp. $R^{(N)}_{k,l}(t)$) be the number of infected (resp. recovered) (k,l) nodes in this network at time t. Then in a small time interval Δ,

$$\begin{aligned} I^{(N)}_{k,l}(t+\Delta) = I^{(N)}_{k,l}(t) &+ \text{number of } (k,l) \text{ nodes infected in time } \Delta \\ &- \text{number of } (k,l) \text{ infected } (k,l) \text{ nodes that recover in time } \Delta. \end{aligned}$$

Since each infected node recovers after an exponentially distributed time of rate ν, the number of (k,l) infected nodes that recover in Δ will be approximately $I^{(N)}_{k,l}(t)\nu\Delta$. There will be additional terms containing Δ^2 which we neglect.

Let us compute the number of (k,l) nodes that get infected in time Δ. There are $N^{(N)}_{k,l} - (I^{(N)}_{k,l}(t) + R^{(N)}_{k,l}(t))$ susceptible (k,l) nodes. Assume that each (k,l) node has a probability $p_{k,l}$ to get infected in the interval Δ. Then, expected number of infected (k,l) nodes in time Δ will be $(N^{(N)}_{k,l} - (I^{(N)}_{k,l} + R^{(N)}_{k,l}))p_{k,l}$.

Each (k,l) node has k incoming edges. Assuming that the edges are connected independently, $p_{k,l} = (1-(1-q_{k,l})^k)$, where $q_{k,l}$ is the probability that the infection is transmitted along one of the edges in Δ. The number of (k,l) nodes infected in Δ is thus

$$(N^{(N)}_{k,l} - (I^{(N)}_{k,l} + R^{(N)}_{k,;})) \cdot (1-(1-q_{k,l})^k),$$

which, for Δ sufficiently small, can be approximated as

$$(N^{(N)}_{k,l} - (I^{(N)}_{k,l} + R^{(N)}_{k,l}))kq_{k,l}.$$

Finally, we compute $q_{k,l}$. In an interval Δ, each infected node transmits the infection with probability $\lambda\Delta$. Thus, there are $\sum_{k',l} l' I^{(N)}_{k',l'}\lambda\Delta$ edges that are infected and that transmit the infection in Δ. There are a total of $\sum_{k',l'} l' N^{(N)}_{k',l'}$. Assuming that an incoming edge is connected uniformly at random to an outgoing edge, we obtain

$$q_{k,l} = \frac{\sum_{k',l'} l' I^{(N)}_{k',l'}\lambda\Delta}{\sum_{k',l'} l' N^{(N)}_{k',l'}}.$$

Consequently,

$$I^{(N)}_{k,l}(t+\Delta) - I^{(N)}_{k,l}(t) = (N^{(N)}_{k,l} - (I^{(N)}_{k,l} + R^{(N)}_{k,l}))k\frac{\sum_{k',l'} l' I^{(N)}_{k',l'}\lambda\Delta}{\sum_{k',l'} l' N^{(N)}_{k',l'}} - I^{(N)}_{k,l}\nu\Delta.$$

If the initial number of nodes is large, then we can divide the two sides of the above equation to obtain the following difference equation in terms of the fraction of the infected and the recovered nodes:

$$i_{k,l}(t+\Delta) - i_{k,l}(t) = (f(k,l) - (i_{k,l}(t) + r_{k,l}(t)))k\frac{\sum_{k',l'} l' i_{k',l'}(t)\lambda\Delta}{\sum_{k',l'} l' f(k',l')} - i_{k,l}(t)\nu\Delta.$$

To complete the picture, we take the limit $\Delta \to 0$, and obtain the differential equations (1) and (2). □

Remark 1. In the above equation $i_{k,l}$ is the fraction of the (k,l) nodes that are infected. This fraction varies between 0 and $f(k,l)$. If instead, we want to look at the evolution of the fraction of infected nodes and recovered nodes conditioned on them being (k,l) nodes, then the corresponding differential equations for these fractions will be

$$\frac{di_{k,l}(t)}{dt} = \lambda k(1 - i_{k,l}(t) - r_{k,l}(t))\frac{\sum_{k',l'} l' f(k'.l') i_{k',l'}(t)}{\sum_{k',l'} l' f(k',l')} - i_{k,l}(t)\nu, \quad (3)$$

$$\frac{dr_{k,l}(t)}{dt} = i_{k,l}(t)\nu. \quad (4)$$

3 Epidemics without recovery

The solution of (1) and (2) can be computed numerically. In some specific case we can obtain explicit solutions to these equations. In particular, this is the case when there is no recovery: $\nu = 0$, or in the language of Twitter, they keep generating new tweets with the same hashtag. That is, a hashtag never gets out of mode. This can well represent the case for the topics or personalities that can sustain popularity over a long period of time.

Since there are no recovered nodes, $r_{k,l}(t) = 0,\ \forall t$, and (1) takes the form

$$\frac{di_{k,l}(t)}{dt} = \lambda k(f(k,l) - i_{k,l}(t))\frac{\sum_{k',l'} l' i_{k',l'}(t)}{\sum_{k',l'} l' f(k',l')}. \quad (5)$$

The differential equation (5) can be solved in terms of a reference value of (k,l), say $(k,l) = (1,1)$ by noting that

$$\frac{1}{k(f(k,l) - i_{k,l}(t))}\frac{di_{k,l}(t)}{dt} = \frac{1}{(f(1,1) - i_{1,1}(t))}\frac{di_{1,1}(t)}{dt},$$

whence

$$f(k,l) - i_{k,l}(t) = \frac{f(k,l) - i_{k,l}(0)}{(f(1,1) - i_{1,1}(0))^k}(f(1,1) - i_{1,1}(t))^k =: c_{k,l}(f(1,1) - i_{1,1}(t))^k. \quad (6)$$

The fraction of infected nodes of degree $(1,1)$ can be obtained by substuting the value of $i_{k,l}(t)$ in (5) and solving it:

$$\frac{di_{1,1}(t)}{dt} = \lambda(f(1,1) - i_{1,1}(t))\frac{\sum_{k',l'} l'\left(f(k',l') - c_{k',l'}(f(1,1) - i_{1,1}(t))^{k'}\right)}{\sum_{k',l'} l' f(k',l')}. \tag{7}$$

Deterministic in-degree

Assume that the in-degree K is deterministic and is equal to d. Then, equation (5) becomes

$$\frac{di_{d,l}(t)}{dt} = \lambda d(f(d,l) - i_{d,l}(t))\sum_{l'} \frac{l' i_{d,l'}(t)}{\sum_j j f(d,j)}.$$

Since the expected in-degree and the expected out-degree coincide, $\sum_j j f(d,j) = d$. Denote $\Theta(t) = \sum_j \frac{j i_{d,j}(t)}{d}$, and rewrite the above equation as:

$$\frac{di_{d,l}}{dt} = \lambda d(f(d,l) - i_{d,l}(t))\Theta(t). \tag{8}$$

Multiplying the above equation by $\frac{l}{d}$ and summing over all values l, we obtain the following equation for Θ:

$$\frac{d\Theta}{dt} = \lambda d(1-\Theta)\Theta,$$

which upon integration yields:

$$i_{d,l}(t) = f(d,l) - c_1 e^{-\lambda d \int \Theta(t)dt} = f(d,l) - \frac{f(d,l) - i_{d,l}(0)}{1 - \Theta(0) + \Theta(0)e^{-\lambda d \cdot t}}. \tag{9}$$

4 Formal justification of convergence to ODE

In this section, we sketch the proof of convergence. We follow the generator approach as expounded in Ethier and Kurtz [9]. To this end, we must first specify in detail the Markov process from which we start. Note that in the spirit of the directed configuration model and the "deferred decision principle" (see e.g. [10], Section 1.3), the network is in fact constructed on the fly. This is done for reasons of tractability. It is mathematically much more difficult to prove the alternative, where at first a random network is generated and then the dissemination is performed on this fixed network. Another simplifying assumption is that we impose an upper bound on the in- and out-degree (say K). It is the subject of future research whether these assumptions can be lifted.

From the description in the previous section, we define for a network with N nodes a Markov process with state space $S_N = \{(i_{1,1}/N, \cdots, i_{K,K}/N, r_{1,1}/N, \cdots, r_{K,K}/N) : i_{k,l} + r_{k,l} \leq f(k,l)\}$ and generator Q_N satisfying:

$$Q_N g(\mathbf{i}, \mathbf{r}) = \nu \sum_{k,l}^{K} N i_{k,l} [g(\mathbf{i} - N^{-1}\mathbf{e}_{k,l}, \mathbf{r} + N^{-1}\mathbf{e}_{k,l}) - g(\mathbf{i}, \mathbf{r})] \tag{10}$$

$$+ \lambda \sum_{k,l}^{K} (N_{k,l}^{(N)} - N i_{k,l} - N r_{k,l}) k \tilde{q}_{k,l} [g(\mathbf{i} + N^{-1}\mathbf{e}_{k,l}, \mathbf{r}) - g(\mathbf{i}, \mathbf{r})], \tag{11}$$

where $\tilde{q}_{k,l} := \Delta^{-1} q_{k,l}$ and $\mathbf{e}_{k,l}$ denotes the (k,l) unit vector.

Using Chapter 8, Coll. 3.2 from [9], we can show that this process is a Feller process. By Taylor-expanding the generator Q_N, we find the following generator to be the candidate limit process:

$$Q = \nu \sum_{k,l}^{K} i_{k,l} [-\partial_{i_{k,l}} + \partial_{r_{k,l}}] + \lambda \sum_{k,l}^{K} (f(k,l) - i_{k,l} - r_{k,l}) k \tilde{q}_{k,l} \partial_{i_{k,l}}, \tag{12}$$

as $N^{-1} N_{k,l}^{(N)} \to f(k,l)$. Note that this corresponds to the deterministic process which satisfies the system of ODEs as in Theorem 1. We must check the conditions of Theorem 6.1 in Chapter 1 of [9], of which the Feller property of the limit process is the only non-trivial part. When the dimension is finite (which is why we bound the maximal degree) we can show this by Chapter 8, Thm. 2.5 from [9].

5 Numerical experiments

In this section, we validate the mean-field model developed in Section 2. First, we generate a Configuration-type graph as was explained in that section. Briefly, the in-degree and out-degree sequences are generated according to the given degree distributions. So as to have the same number of incoming stubs as outgoing stubs, the difference between the two is added to the smaller quantity. A graph is then created by matching an incoming stub with an outgoing stub chosen uniformly at random.

For computing the solution of the system of differential equations (1) and (2) numerically, the empirical degree distributions from the graph generated previously are given as input.

The results of two such experiments with 20000 nodes is shown in Figure 1. The in-degree and the out-degree sequences were taken to be independent of each other. The out-degree sequence was drawn from a Uniform distribution in the set $\{1, 20\}$ in the two simulations. For the figure on the left, the in-degree distribution was taken to be deterministic with parameter 10, and for the figure on the right it was the Zipf law on $\{1, 71\}$ and exponent 1.2. In both experiments, $\lambda = 1$ and $\nu = 0.5$, and 5 percent of all nodes were assumed to be infected at time 0.

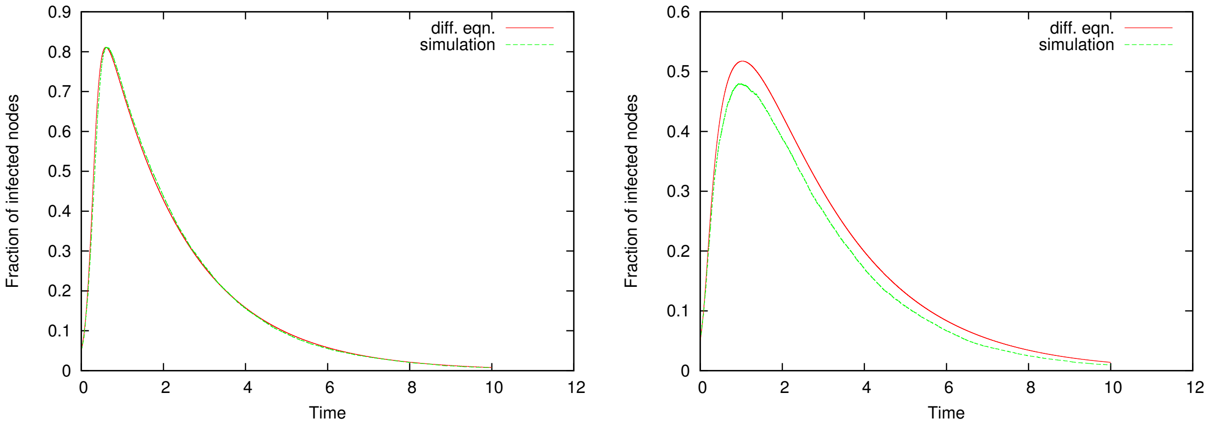


Fig. 1. Fraction of all nodes infected as a function of time for Deterministic in-degree distribution (left) and Zipf in-degree distribution (right).

It is observed that the dissemination process is faster when the variance of the in-degree distribution is smaller. This observation was reinforced by other experiments in which the in-degree was drawn from a Uniform distribution. In several other experiments that we conducted, it was also observed that the out-degree distribution does not have any noticeable effect of the dynamics of the epidemics.

6 Summary and future work

We proposed mean-field differential equations to model the lifecycle of hashtags in directed social networks. In networks with large number of nodes, these equations model reasonably well the dynamics obtained from simulations. It was observed that the variance of the in-degree distribution has a negative influence on the speed of the information dissemination, that is, higher the variance of the in-degree distribution, slower is the rate of dissemination. On the other hand, the dynamics are insensitive to the out-degree distribution.

There are several assumptions inherent in the model that need to be validated in real social networks. Firstly, the assumption of exponentially distributed infection and recovery rates need not hold in practice. For distributions other than the exponential distribution, the derivation and the justification of the mean-field model is more technically involved. Secondly, we assume that the Configuration-type graph is a good model for social networks.

Our on-going work is oriented towards investigating the influence of the variance of the in-degree distribution and giving a theoretical foundation to the above observations. We also intend to compare the dynamics computed using the mean-field model and those obtained from the Twitter graph to check validity of the proposed model and make improvement as necessary.

References

1. Chen, N., Olvera-Cravioto, M.: Directed random graphs with given degree distributions. To appear in Stochastic Systems
2. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Physical review letters **86**(14) (2001) 3200
3. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. The European Physical Journal B-Condensed Matter and Complex Systems **26**(4) (2002) 521–529
4. Ganesh, A., Massoulié, L., Towsley, D.: The effect of network topology on the spread of epidemics. In: INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. Volume 2., IEEE (2005) 1455–1466
5. Van Mieghem, P., Omic, J., Kooij, R.: Virus spread in networks. Networking, IEEE/ACM Transactions on **17**(1) (2009) 1–14
6. Durrett, R.: Random Graph Dynamics. Volume 20. Cambridge university press (2007)
7. Barrat, A., Barthelemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press (2008)
8. Draief, M., Massouli, L.: Epidemics and Rumours in Complex Networks. Cambridge University Press (2010)
9. Ethier, S., Kurtz, T.: Markov processes: Characterization and convergence. Wiley Interscience (2005)
10. Mitzenmacher, M., Upfal, E.: Probability and Computing : Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, 2005.

Spatial Explicit Model to Visualize the Spread of Epidemic Disease in a Network

Mohan Timilsina[1], Raphael Duboz[1] and Hideaki Takeda[2]

[1] Computer Science and Information Management,
Asian Institute of Technology,
Pathumthani, Thailand
http://www.ait.ac.th

[2] National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda,
Tokyo, Japan
http://www.nii.ac.jp

Abstract. The development of the visualization tool of any infectious spreading disease requires an elegant graphical user interface which can show the result of simulation at every time steps. The goal of this work is to provide an effective way to visualize epidemic diseases, e.g., flu, cholera, influenza, etc. in a network for the decision makers to determine the level of epidemic threats in any vulnerable zones. The Gravity and Susceptible, Infected and Recovered (SIR) epidemic models are used to draw a network on the map and simulate the disease status. Thus the results obtained are visualized to give a clear picture of spreading behavior of a disease. This visualization tool provides an understanding for the epidemical analysis of diseases through graphs and network statistics.

Key words: Susceptible Infected Recovered, gravity model, visualization tools, network

1 Introduction

The epidemic of any infectious disease can be better understood by decision makers if they are provided with a proper tool to visualize this. If we consider an example for a new outbreak of infectious disease in a particular geographical region it is always crucial for a decision maker to come out with epidemic scenarios. In fact, visualization of the vulnerable zone plays an important role for understanding populations from serious epidemic hazards. As a result of simulation of diseases in a network, early visualization of such threats can be obtained. Hence, a well-developed decision support system with simulations to identify the vulnerable zone has become a common area of interest in this field. The growing interest in simulation modeling techniquea has been used in the context of visualizing epidemic diseases. A review of some existing epidemic disease visualization tools using different technologies follows. GLEAMviz [1] visualization software has been made to visualize epidemic levels on a worldwide scale. This project uses the multi scale mobility model from the airport listed in International Air Transport Associations (IATA) database for the disease transmission route. Geographical Information Systems (GIS) and electronic disease surveillance databases [2] are extensively used to visualize the impact of epidemics of geographical regions. Furthermore, social network analysis [3] is also used for epidemic simulations for disease transmission in human beings. Although the above mentioned technologies provide epidemic visualization they need huge data sets about diseases and are not flexible to be applied in different emerging contagious diseases because they rely on epidemic network data sets [4]. Hence a decision support system that can provide policy makers with the ability to create epidemic scenarios in absence of disease routes for an emerging contagious disease seems to be essential. To address this problem an epidemic visualization system with simulation is proposed. The purpose of this tool is to visualize epidemic disease networks through simulation results. The specific goals are listed as follows:

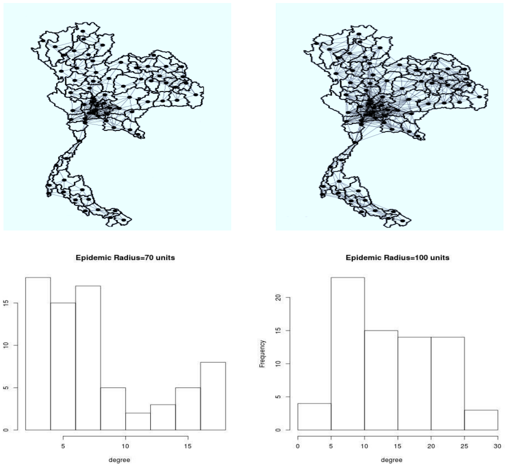
- To use gravity model for epidemic disease network formation;
- To use SIR for epidemic modeling; and
- To visualize epidemic scenarios on a map.

The epidemic disease network is formed by using the gravity model. This model provides the rate of disease transmission across distance. The strength of disease transmission is decreased with the increase in distance Zipf in the 1940's [5] has provided a theoretical motivation for movement between cities 1 and 2 being governed by a $\frac{P_1P_2}{d}$ relationship, where P is the respective city population and d is separation distance. The strength of the movement is affected by the size of the population and their separations. Hence, in our tool we used distance as an important factor to construct the network. The closer the nodes the denser is the network. The visualization of epidemic diseases for decision support system using Susceptible(S), Infected(I) and Recovered(R) (SIR) epidemic model is one of the main components of this tool. It is a good modeling tool for many infectious disease like small pox [6], influenza and measles, etc.; where every population can be categorized into three different stages [7].It has three stages and in a population which are prone to disease are in Susceptible stage. Those who carry the disease and transmit the disease to others are in Infected stage and those who are immune or recovered from the disease and those removed from the population are in Recovered stage. In the tool the three stages of the model are depicted in the nodes which in this case are cities on the map. This tool is intended for general purpose epidemic disease visualization to decision makers at different disease transmission rates, epidemic distances,simulation timesteps and infection probabilities. The results thus obtained are visualized on a map. The system uses two dimensional coordinates of cities in the map to draw a network underlying the principle of gravity model [8].The system is very flexible to change the network as per the necessity of decision makers, whereas previously developed technologies requires data as a fixed source for the network formation. Section 2 introduces network topology. Section 3 is about the SIR epidemic model, Section 4 provides the system description, Section 5 describes the implementation of the system, Section 6 describes simulation and analysis, Section 7 is about user study and evaluation, Section 8 discusses about the limitations of the system and Section 9 draws conclusions.

2 Network Topology

The strength of interconnection of two nodes is determined by how closely they are located. The affiliation between the cities is the function of the distance and thus the network is built upon physical proximity between them. The force of epidemics decays with the increase in distance [9].The spread of any infectious disease can be depicted by human travel pattern between two cities. An analogy with the physics law, it is expected that disease transmission increases with decrease in geographical distance [10]. The generic model for the gravity is described as:

$$F_{i,j} = G\frac{M_iM_j}{d_{i,j}^2}$$



(a) (b)

Fig. 1. Example of epidemic network . (a) Networks with 70 units epidemic radius. (b) Networks with 100 units epidemic radius.

where G is the gravitational constant: M is the mass of the bodies: i,j and d is the distance between them. In the context of this work the force of an epidemic is taken as the link formation between the two nodes, i.e., two nodes will construct the link if they fall into the range of the epidemic distance. This means connectivity completely depends upon closeness of the cities. Fig. 1 visualizes the network of cities with two different epidemic radius and its degree distribution.

3 Epidemic Model

The core of the system is based upon epidemiological states of the cities. Hence the connection among different cities is a vital part of network formation. The transportation network is also an important role for the disease carrier [11].Every connected city is susceptible to the disease. Those cities which are already infected will either stay infected or will change the status to be recovered. The spreading process of disease can be understood as: at each time step the susceptible nodes becomes infected at a rate of β; the infected nodes will be recovered at a rate of γ. The transmission of the disease in the nodes is $S \xrightarrow{\beta} I \xrightarrow{\gamma} R$. SIR model is represented by the differential equation as [12][13]

$$\frac{ds}{dt} = -\beta is, \frac{di}{dt} = \beta is - \gamma i, \frac{dr}{dt} = \gamma i$$

Each state is fraction of population at a particular timestamp. The population is one of the stages at a particular time. This model is particularly meant for fast

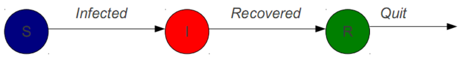


Fig. 2. SIR epidemic model.

spreading disease and for a constant population. Fig. 2 visualizes the mode of spread of disease stages.
The overall composition of the cities in a map will be of three different kinds of categories: susceptible cities, infected cities and recovered or removed cities.

4 System Description

4.1 Methodology

The system is developed by integrating the Geographical Information System (GIS) shape files for reading the map, graph theory for analyzing the network and SIR epidemic model for the simulation. The input of the system is epidemic radius, disease transmission rate , simulation timestep and infection period. The computation of the network data is done by the open source boost library[3] and for the visualization open source 2D Cairo graphics[4] is used. The system first reads the GPS position of the cities and locates on a map. Then after getting the input for the epidemic radius it draws the network on top of it. The user can then select any city on a map as an epidemic outbreak and provide the transmission probability[14] and simulation timestep and can see the epidemic visualization on a map.

4.2 Simulation Model

In the SIR epidemic model the cities on the map are classified according to their states: susceptible(S), infected(I) and recovered(R).The epidemic spreads through the links between the cities. The spreading process of disease can be understood as: At time (t=0) the node infects its neighboring susceptible nodes with the random probability(p) less than the transmission probability(tp). At time (t=1) the neighbor node gets infection and at time (t=2) the initial node which has transmitted the disease passes the infectious period and becomes immune or recovered in the system. The overall scenario can be viewed in the figure below:

[3] http://www.boost.org/
[4] http://www.cairographics.org/

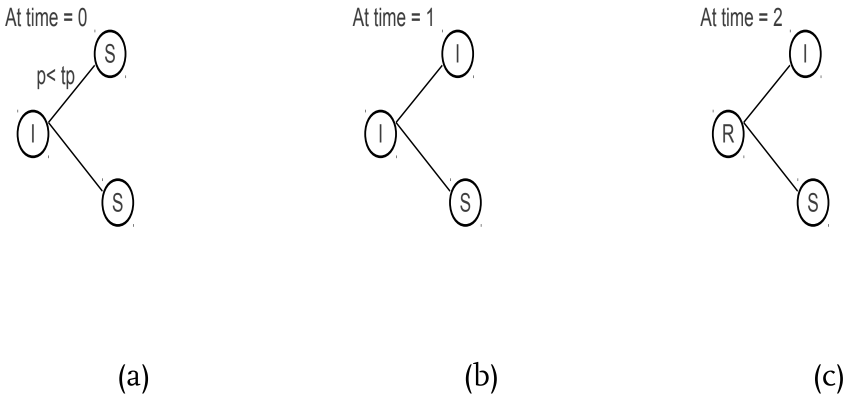


Fig. 3. S,I and R are the Susceptible, Infected and Recovered nodes of the disease network. (a) Infected nodes transmits the disease. (b) Susceptible nodes receive infection. (c) Recovered nodes recover from infection.

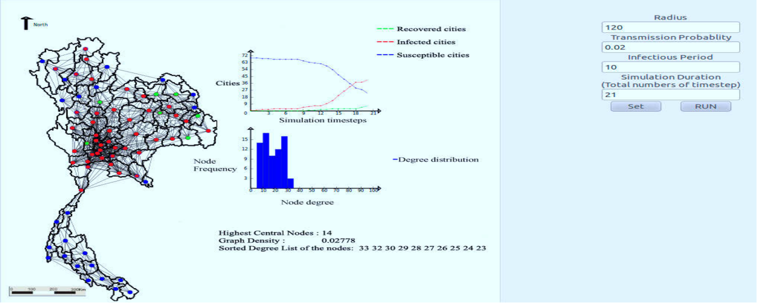
Pseudocode for the infection and recovered status

```
program Infective and Recovered nodes (Output)
  {Get the transmission probablity,infectious period from a user
   For every nodes which is infected state:=1};
   var
    nodes: 0..maxNodes;
    statusVector;
    timeVector;
    transmissionProbablity,randomProbablity: Real;
   begin
    nodes := 0;
    statusVector:= 0;
    timeVector  := 0;
   repeat
    randomProbablity:= rand() / (RAND_MAX+1))*-1;
    if(randomProbablity<transmissionProbablity)
     {
     statusVector:=1; //infects the randomly selected outdegree nodes
       timeVector :=timeVector+1;
     }
    if(timeVector>infectiousPeriod &&  statusVector==1)
     {
       statusVector:=2; //Nodes will be recovered or removed
     }
   nodes:= nodes+1;
   until nodes = maxNodes;
end.
```

The model works undertaking two major widely used components in epidemiology called as infectious period to produce the symptoms of disease in the nodes and transmission probability to transfer the disease from infected to the susceptible nodes. Thus the epidemic visualization system, that enhanced by this model, will have sufficient information to simulate the disease and create epidemic scenario in a disease network.

5 Implementation

The proposed system is made which allows the user to enter the epidemic radius, transmission probability, infectious period of the disease and simulation time steps. The network is then drawn by calculating the Euclidean distance between the pairs of the cities. From,the whole cities user can select any of them as an outbreak of the epidemic and run the simulations and the result is dynamically visualized on the map for every timestep. At the same time the user can also see the graph density, degree distribution and epidemic curves in the interface. Figure below is the User interface of the prototype.

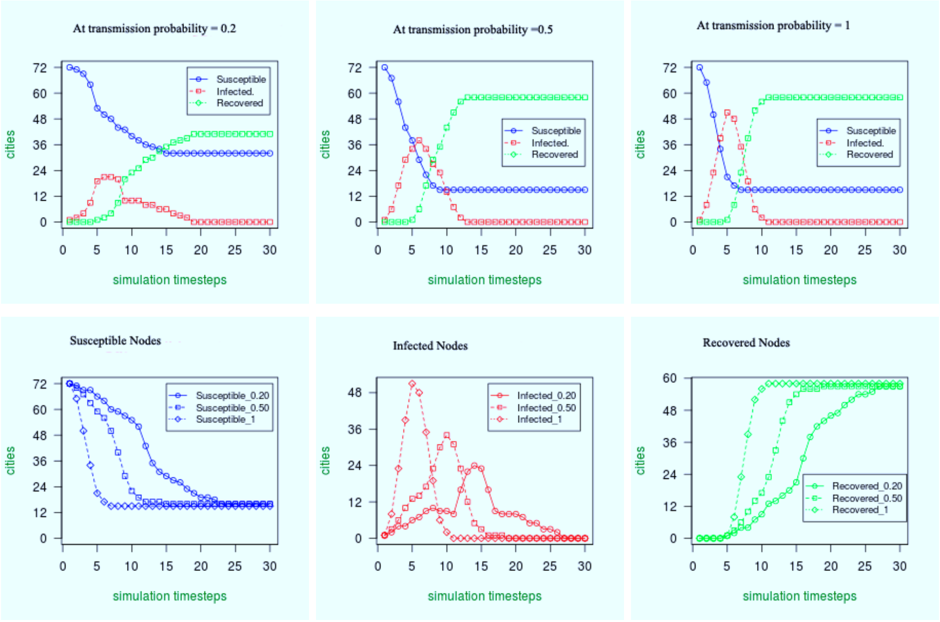


Highest Central Nodes : 14
Graph Density : 0.02778
Sorted Degree List of the nodes: 33 32 30 29 28 27 26 25 24 23

Radius
120
Transmission Probablity
0.02
Infectious Period
10
Simulation Duration
(Total numbers of timestep)
21

Fig. 4. Screenshot of the epidemic disease visualization of the proposed system.

6 Simulations and Analysis



(a) (b) (c)

Fig. 5. SIR Simulation result for 60 unit epidemic radius

In figure(a), Susceptible curve decreases rapidly with the increase in transmission probability, with the increase in simulation timesteps the curve starts to decrease and after some period the curve flattens out. In figure(b), Infected curve increases rapidly with the increase in transmission probability and reaches the peak of the infection, as the number of simulation timeteps increases the infection curve also falls down rapidly. It starts to decrease and finally the curve reaches to minimum 0 and flattens out where no infection is seen. Similarly in figure(c) the recovery node is only seen after the infectious period of the outbreak in this case the infectious period of the disease is four days. It is observed from the figure(c) that the recovery curve also increases rapidly with the increase in transmission probability and after some period the curve reaches to the peak and becomes constant.

7 User Study and Evaluation

Users of this prototype system are only few members including the developer and we have not verified results yet. However we think this prototype will be able to create the epidemic scenarios for decision makers without explicitly requiring the network construction data. Apart from the user study the evaluation of this prototype has been done with the existing visualization tool Epigrass[5] and Gleamviz[6]. The results of the evaluation follows:

1. **Ability to visualize SIR epidemic disease:** The developed prototype fully supports this feature by asking the user to set up their parameters in the tools. GleamViz and Epigrass can also support this feature moreover they can be modeled for all types of epidemic disease.
2. **Ability to visualize the disease in a local scale:** The developed prototype and Epigrass fully supports this feature by prompting the user to load their specific GIS shape files and to run the simulations whereas GleamViz provides the global level epidemic visualization.
3. **Ability to create epidemic scenarios:** All of these tools support this feature.
4. **Ability to create the disease network with the Euclidean distance between the nodes:** This is an interesting feature of the prototype where the network is drawn underlying straight line distance between the two nodes. Due to this reason the prototype does not require the network construction data else it only needs the epidemic radius from the user. Whereas Gleamviz requires multi scale mobility model of the International Air Transport Association (IATA) database for the network formation and Epigrass uses the transportation and passenger flow data for the network formation. To the best of our knowledge in the absence of the network data in GleamViz and Epigrass both cannot simulate and visualize the epidemic disease.

7.1 Comparision of the prototype with the existing visualization tools

Table 1 provides a comparision of the prototype with the GLEAMViz and EpiGrass visualization tools.

[5] http://sourceforge.net/projects/epigrass/

[6] http://www.gleamviz.org/

Table 1.

Criteria	*Prototype*	*GLEAMviz*	*EpiGrass*
Visualization of spatial data	Yes	Yes	Yes
Visualization of models	Yes	Yes	Yes
Visualization of simulated results	Yes	Yes	Yes
Types of epidemic disease	Contagious acute infectious disease	Every kind of epidemic disease	Every kind of epidemic disease
Data for disease network formation	No network data requirement	IATA databases	Transportation and passenger flow data
Software modularity and extensibility	Yes	Client server based online tool	Yes

The strength of the prototype is to draw a network for the disease simulation in the absence of the network data which is not present in either of the two systems. The prototype calculates the Euclidean distance between the nodes within the epidemic radius and draws the network on the map. Currently the prototype is only applicable for the visualization of acute infectious disease which has a very short exposed period. Table 1 compares the important feature of the popular epidemic visualization tools with the prototype.

8 Limitations

The prototype is developed using SIR epidemic model. This model also has shortcomings because it does not consider the immigrants and emmigrants of the population and assumes the total population as constant [15]. In the prototype we have cities and assume the population is well mixed and the same in all parts of the country which is completely impossible in a real situation. In the developed prototype the latent period for acquisition of infection and the start of infectiousness has not been addressed due to which the chronic infectious disease cannot be modeled. The diseases with latent periods that can be modeled using the extensible SIR model with latent period called SEIR epidemic model cannot be used by this prototype. However the prototype can be useful for the decision makers to observe and create the scenarios for the acute infectious disease having a very short exposed period like influenza, chicken pox, distemper. etc. In the current version of the system we can only address the SIR model but in our upcoming version of the system we will also be able to provide the flexiblility to change the infectious disease model like SEIR, MSIR, SIS and SIQR to visualize the epidemic spread of the disease.

9 Conclusion and Future Work

This research proposes the visualization of the epidemic disease using a simple epidemic and gravitational model. To accomplish this a prototype is developed without explicitly requiring predefined network construction data. The epidemic spread as a result of simulation is visualized in the map. The usage of graph theory has become very useful to visualize the epidemic network, degree distribution and finding the central nodes in the network. The developed prototype can visualize the states of every node at the given time stamp. This tool is very flexible for visualizing SIR modeled infectious disease for a given transmission rate and epidemic radius. Currently, this prototype cannot address all the epidemic diseases. In the current context we also lack the real epidemic data to test with our system generated results. Due to this reason we cannot ensure that the prototype can be very accurate with its simulated results. However in the upcoming version the prototype will come out with tested real data and also be capable of visualization for other existing epidemic models. Therefore the proposed visualization approach is enhanced in the further development of the tool.

References

1. Broeck, W.V.d., Gioannini, C., Goncalves, B., Quaggiotto, M.,Colizza, V., : The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale J. BMC. Infectious.Diseases. 11,37 (2011)
2. Burns, J., Hatt, C., Brooks, C., Keefauver, E., Wells, E.V., Shuchman, R., Wilson, M.L.,: Visualization and Simulation of Disease Outbreaks: Spatially-Explicit Applications Using Disease Surveillance Data. In: 26th Annual Esri International User Conference, Redlands,CA (2006)
3. Bisong, H., Jianhua, G.,: Simulation of Epidemic Spread in Social Network. In: Management and Service Science, International Conference , Wuhan, China (2009)

4. Chen, H., Zeng, D.,: AI for Global Disease Surveillance J. Intelligent Systems, IEEE. 24, no.6, 66-82(2009)
5. Zipf GK.,: The P1 P2/D Hypothesis: On the Intercity Movement of Persons. Am Sociol Rev 11: 677-686 (1946)
6. Ferguson, N. M., Keeling, M. J., Edmunds, W. J., Gani, R., Grenfell, B. T., Anderson, R. M., Leach, S.,: Planning for smallpox outbreaks. Nature, 425(6959), 681-685 (2003).
7. Antulov, N., Lančić, A., Štefančić, H., Šikić, M.,: FastSIR algorithm: A fast algorithm for the simulation of the epidemic spread in large networks by using the susceptible-infected-recovered compartment model, Inf. Sci. 239 226-240(August 2013)
8. Eggo, M.R., Cauchemez, S., Ferguson, N.M.,: Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States J. J.R.Soc.Interface, 8, 233-243(2011)
9. Barthélemy, M.,: Spatial networks, Physics Reports, 499:1–101,(2011)
10. Sarzynska, M., Udiani, O., Zhang, N.,: A study of gravity-linked metapopulation models for the spatial spread of dengue fever, CORD Conference Proceedings (August 2013)
11. Zhang, Y., Zhang, Yo., Liu, Z.,: The role of different transportation in the spreading of new pandemic influenza in mainland China. In: Geoinformatics, International Conference , Beijing, China (2011)
12. Newman, M. E.: Spread of epidemic disease on networks. Physical review E, 66(1), 016128 (2002).
13. Hethcote, H. W. :The mathematics of infectious diseases. SIAM review, 42(4), 599-653 (2000)
14. Meyers, L. A., Pourbohloul, B., Newman, M. E., Skowronski, D. M., Brunham, R. C.,: Network theory and SARS: predicting outbreak diversity. Journal of theoretical biology, 232(1), 71-81 (2005).
15. Ahmed, E., Agiza, H. N.,:On modeling epidemics including latency, incubation and variable susceptibility. Physica A: Statistical Mechanics and its Applications, 253(1), 347-352 (1998).

Predicting Network Structure Using Unlabeled Interaction Information

Mehwish Nasim and Ulrik Brandes

Department of Computer and Information Science
University of Konstanz
{mehwish.nasim,ulrik.brandes}@uni-konstanz.de

Abstract. We are interested in the question whether interactions in online social networks (OSNs) can serve as a proxy for more persistent social relation. With Facebook as the context of our analysis, we look at commenting on wall posts as a form of interaction, and friendship ties as social relations. Findings from a pretest suggest that others' joint commenting patterns on someone's status posts are indeed indicative of friendship ties between them, independent of the contents. This would have implications for the effectiveness of privacy settings.

Key words: Link inference, Facebook, Interaction

1 Introduction

Sociological research has identified various dimensions of social relations (e.g., time, affect, intimacy, or reciprocal services [1]) and group formation (e.g., shared interests, personal preferences, ascribed status [2]). Closeness in these dimensions brings individuals together because they organize their relations around common *foci* [3]. A focus is defined as a social, psychological, legal or physical entity around which joint activities take place, e.g., offices, voluntary communities, hangouts, families, etc. Applying focus theory to online social networks (OSNs), we should therefore expect that users who are friends with each other share similar interests and behave accordingly. For instance they might join similar groups (on Flickr, Facebook etc.), "like" similar pages (e.g., on Facebook), "follow" similar accounts (on Facebook, Google+, Twitter etc.), or engage in similar discussions.

The relation between network structures and interaction patterns of online users is an active area of research. Various approaches have been devised to predict links in social networks [4,5,6]. Studies have focused on properties such as (partially) known network structures, actor attributes, and interaction patterns to deduce further information. Horvát *et al.* [7], for instance, show that the combination of knowledge of confirmed contacts between members on a social network and their email contacts to non-members provides enough information to deduce a substantial proportion of the relationships between non-members. Recently, Romero *et al.* [8] studied the interplay between network structures and topical interests on Twitter, and were able to predict links between users on Twitter from topical affiliations inferred from their usage of hashtags. Hashtags were used to define user sets, which can be viewed as embedding Twitter users in the topics associated with those hashtags.

In a formative study we investigate whether interactions can provide information on network ties even without content knowledge, i.e., instead of topical groups as defined, say, by hashtag usage, we make use of participation in discussions only, irrespective of their focus. Using a Facebook dataset collected in a previous study [9], we find that simple discussion features provide sufficient information to reconstruct friendship networks to a large degree. If this holds more generally, it implies that hiding one's friendship is not as effective as one might hope because it is undermined by one's interaction behavior.

2 Dataset

The context of our analysis is Facebook, a social networking site where users can add others as friends to their profiles. In addition, Facebook provides various interaction options to its users. One of the ways in which friends (the alters) interact with a focal user (the ego) is by posting comments on the discussions (called status posts) started by that user. In the scope of this paper, we are analyzing the commenting history of the alters on status posts made by the ego on what is called his or her "wall."

For a plausibility test of the hypothesis that interactions are closely related to ties, we use data originally collected to study commenting behavior of Facebook users [9]. This study found that the likelihood of commenting on a post increases in the presence of previous comments from members of the same community of friends, where communities are determined by partitioning the friends of a user based on their mutual friendship ties. The dataset consists of 50 Facebook profiles

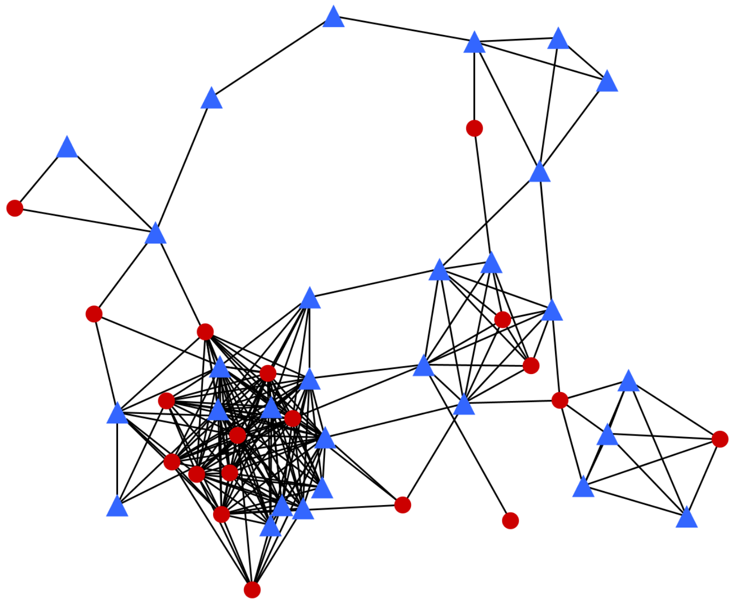


Fig. 1. Personal network comprised by the Facebook friends of a profile and their friendship ties. Alters who took part in discussions are shown as blue triangles. Note that the focal profile (i.e., the ego) would be connected to all others and is therefore omitted.

for which a total of $5,778$ status updates and $40,186$ comments have been recorded over a three-day period. On average, users in this data set had 170 friends and posted 116 status updates for an average of 7 comments.

No topical or otherwise identifying information was recorded for status posts. The only information available is the identity of posters and commentors. For the rest of this paper we collectively refer to a status post and comments on that post as a discussion.

3 Inferring Friendship Ties

Our goal is to infer friendship ties from participation in discussions. Suppose that we have the discussions related to a Facebook ego profile available to us but we do not have access to any information about the network structure, i.e., no friendship ties between alters are known to us. We will use the following notation:

- $u_1, ..., u_n$ denotes the n alters in an ego's personal network.
- $d_1, ..., d_m$ denotes the m discussions in a profile.
- $D(u_i)$ denotes the set of discussions in which u_i made a comment.
- $U(d_j)$ denotes the set of users who commented in discussion d_j.

Very simple features are extracted from the discussions to predict the presence of friendship relation between pairs of alters. To determine the behavioral similarity of two alters based on the discussions they are part of. Recall that the discussions are not labeled unlike hashtags in Twitter where each hashtag defines a topic of the conversation. While, for example, *#socnet2014* may be a hashtag for discussion on SOCNET 2014, an ego may make several posts related to SOCNET 2014 on his or her Facebook wall leading to several distinct discussions. This implies that we are underutilizing the information available in principle.

Only the follwing features of discussions are used:

- $|D(u_i) \cap D(u_j)|$,
  the number of common joint participations of u_i and u_j.
- $min_{d \in D(u_i) \cap D(u_j)} |U(d)|$,
  the smallest size of a discussion group containing u_i and u_j.
- $max_{d \in D(u_i) \cap D(u_j)} |U(d)|$,
  the largest size of a discussion group containing u_i and u_j.
- $\frac{|D(u_i) \cap D(u_j)|}{|D(u_i) \cup D(u_j)|}$,
  the similarity of participation in discussion (Jaccard coefficient).

From these features we would like to predict the existence of friendship ties. Note that the data set does contain the actual friendship ties among alters of each ego, except for a few missing ones due to privacy settings. We thus picked one profile (shown in Fig. 1) for which the missing links could be added based on an interview with ego, and trained a simple regression model. Only pairs of users who have commented on the same post at least once are considered, and since there is a high degree of imbalance (with many more absent friendship ties than present ones) we randomly selected an equal number of adjacent and non-adjacent dyads for training.

For comparison, we also trained the model from [7], in which structural features computed on an induced subgraph (representing members of an OSN) are used to predict links between the other nodes (the non-members).

4 Results

For the brevity of discussion we motivate the features and explain the results on the example shown in Fig. 1. It has been used for training the model once as it was recorded and once with the missing links introduced with the help of ego. The results are reported in Tabs. 1 and 2.

Table 1. Prediction accuracies after training on recorded data

Features	Min accuracy	Avg. accuracy	Max. accuracy
Number of common discussions	36.68%	55.92%	72.58%
Size of smallest discussion group	40.06%	54.79%	73.75%
Size of largest discussion group	34.29%	52.63%	70.08%
Jaccard coefficient	35.25%	53.02%	78.31%
All features	49.67%	62.17%	80.62%

Table 2. Prediction accuracies after training on interview-corrected data

Features	Min. accuracy	Avg. accuracy	Max. accuracy
Number of common discussions	49.12 %	58.30%	79.69%
Size of smallest discussion group	47.92%	58.74%	74.01%
Size of largest discussion group	46.57%	56.38%	71.56%
Jaccard coefficient	57.91%	61.45%	84.53%
All features	54.61%	68.92%	82.23%

One intuitive way to measure the similarity between alters is to find the number of common discussion they participated in. However, this measure does not differentiate between the discussions which are of broad interest and those which are of interest to people who share a common foci. The classification based on this feature gives an accuracy of 55.92%. Another way to measure the exclusiveness of the discussions is to consider the minimum/maximum number of participants in a discussion. Nevertheless, these features do not differentiate between alters who always share exclusive discussions versus the alters who less frequently participate in exclusive discussions. We also use Jaccard coefficient as a discussion feature. The prediction accuracy with Jaccard coefficient is anticipated to be higher, however the table shows a lower prediction accuracy on average. When we added the missing links in the data for training the algorithm, the accuracy of Jaccard coefficient significantly improved. The combined accuracy of discussions' features is 62.17%. The prediction accuracy shows both the correct number of 1s and 0s predicted. The p-values we obtained for the logistic regression show that the first three features are statistically significant. The estimate coefficient in the case of number of common discussions is positive, which means higher the number of common discussions, the more is the probability that two individuals have a friendship tie. The estimate coefficients for the size of smallest common discussion group and the size of largest common discussion group are negative. This signifies that smaller the size of a discussion group, the more exclusive it would be which is indicative of the fact that participants involved in the discussion share friendship ties.

As mentioned earlier the data set had some missing friendship links; we mitigated the problem of false negatives by filling in the missing links between alters in order to train our algorithm on correct data. The training profile had six missing links between alters. We then analyzed the amount of network information reflected in the discussions. The prediction accuracies are summarized in Tab. 2. Using all the discussion features, on average we get a prediction accuracy of 68.92% and in some cases it can be as good as 82.23% . Adding the information about the missing links in our training/testing set also shows an improvement in the prediction accuracy of the Jaccard coefficient feature. This feature signifies the similarity between two alters and shows how exclusively they comment together over their entire commenting history.

Romero *et al.* [8], used features of common hashtags to predict links between Twitter users. They considered the follow edges and @edges and used the common hashtags and size of minimum/maximum common hashtags as their features, in addition to using the aggregated features of hashtags. In contrast to using discussions sizes as the aggregate overlap of discussions, we find it prudent to use the Jaccard coefficient to measure the similarity between interaction history of pairs of alters in the context of Facebook. For the follow edges the prediction accuracy of [8] is 74% as compared to the 68% accuracy we are getting for the undirected friendship edges on Facebook. Recall, however, that our discussions are not classified into topics, unlike the hashtags in Twitter. This makes the prediction task more difficult. The @edges require knowledge about the content of the tweets and in our case the content of the discussions. Moreover, they also show a boost in prediction accuracy when information about network edges (other than the ones connected to the two users being considered) is introduced.

In order to compare our work with another state-of-the-art link prediction approach, we also implemented the algorithm of Horvát *et al.* [7] which is based on structural properties of the network. The basic approach classifies nodes into two categories, social networking site members and non-members. In our experiments, the selection of members is based on independent decisions modeled by the random selection of a set of members; the value of parameter ρ is 0.5 (i.e., 50%

of the nodes are members) and α is 0.8 (i.e., 80% of the members have shared information about the edges). The structural features mentioned in the study work well when the features are constructed on erroneous data (i.e., data with few missing links), since the predicted variable may also have several missing links. In our experiments we constructed the structural features on erroneous data, but the predicted variable was inferred from the correct data. Our experiments show that approach used by Horvát *et al.*, gives a prediction accuracy of 65.21%. We then used the interaction features along with the structural features and we get a prediction accuracy of 80.43%. Features based on structural properties of graph are incapable of capturing links where not only the two nodes under consideration have their friendship lists hidden but also the friendship lists of their neighbors are hidden. In such scenarios interaction information acts as a proxy to friendship links.

5 Conclusions

Our preliminary results on re-used data suggest that friendship ties can be inferred from participation in unlabeled discussions using very simple features. The approach is based on the surmise that discussions take place among users with common foci. We used Facebook status posts to predict the links between alters. Using features from unlabeled discussions we get a prediction accuracy of almost 69%. Interaction information appears to be capable of detecting ties in the absence of network data, and improve accracy in the case of partially missing data. We intend to refine our approach to improve prediction accuracy, and to test it in a more elaborate empirical study.

References

1. M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
2. Ronald L Breiger. The duality of persons and groups. *Social forces*, 53(2):181–190, 1974.
3. Scott L Feld. The focused organization of social ties. *American journal of sociology*, pages 1015–1035, 1981.
4. Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

5. David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.
6. Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.
7. Emőke-Ágnes Horvát, Michael Hanselmann, Fred A. Hamprecht, and Katharina A. Zweig. One plus one makes three (for social networks). *PloS one*, 7(4):e34740, 2012.
8. Daniel M Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.
9. Mehwish Nasim, Muhammad U Ilyas, Aimal Rextin, and Nazish Nasim. On commenting behavior of Facebook users. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 179–183. ACM, 2013.

The Marriage Matching Problem with Scope Limited By Social Networks and Limited Time

Debra Hevenstone

University of Bern & University of Zurich

Abstract. This paper looks at the problem of bipartite matching problems using the Gale-Shapely matching algorithm. To date the literature has largely ignored the fact that real-world matching often occurs with limited information. Here we explore what happens to classic matching algorithms when time and search scope are limited. To do so, we employ a simulation, or agent-based model with variably limited search scope and time. Findings suggest that limiting scope via an imposed network (generated using orthogonal rules to match criterion) partially compensates for losses in match utility under limited match time; i.e. limited scope can compensate for limited time. Simulation results also suggest that higher quality or more attractive individuals benefit in terms of match utility from a broader scope, but not from more time. Finally, results suggest that the importance of an individual's own qualities for achieving a high-utility match might be greater for women.

Key words: Matching Algorithms, Bipartite Networks, Simulation, Marriage

1 Introduction

This paper compares the standard Gale Shapely marriage matching algorithm with full information and extensive matching opportunities with search scope limited by social networks and with limited matching opportunities. Results illustrate that given limited matching opportunities, matches with search scope limited by social networks produce better matches, despite the fact that social network generation and match assessment use orthogonal criterion. With unlimited matching opportunities higher utilities are achieved with full-information searches.

2 Background and Theory

There are various different matching problems. Perhaps the most well-known is the "marriage problem" as it is conceived of as a set of men who have ranked preferences if women, and women who have preferences over men, who should be paired such that there are no two people of the opposite sex who would both rather have each other than their current partners (i.e. a "stable" match). There are often multiple stable solutions for a given matching problem. The most well known method to find a stable solution is the "Gale-Shapely Algorithm" [1] in which there are multiple rounds in which unengaged men can propose to their most-preferred woman to whom they have not yet proposed. Women can provisionally accept a suitor, if he is preferred to the current one, but if another, better, suitor makes a later offer, women may reject the provisional suitor for the new one.

Matching algorithms are used in two distinct ways. First they can be used by central clearing houses to find stable matches for real world problems. Matching algorithms are used as such to match medical graduates with residencies, high school students and schools, organ donors and recipients, and law school graduates and firms, to name a few. Second, matching algorithms are used to *model* matching processes that occur without central coordination (e.g. dating, job search, or housing search). One of the key limitations of using these algorithms to model real-world processes is that the algorithm assumes full ranking across all potential matches and extensive matching time ($n2-n+1$ rounds, where there are n men and n women, using the Gale-Shapely algorithm) both of which are not the case in real-world matches that occur without a central clearing house. For example, in the case of romantic matches, there is evidence that in on-line dating (what we might consider a full-information context) yields less similar matches than off-line, likely because off-line people meet one another through one's friends, who tend to be similar [2]. As such, it is unrealistic to assume that individuals can rank the full pool of potential partners or that they have sufficient time to make enough offers to reach the stable solution.

Research in both computer science and economics study matching processes with limited time and/or search scope, varying in how they assume matching might diverge from the ideal conditions assumed in the classic algorithms, and in the consequences they examine, resulting from these limitations (e.g. time to convergence or match stability in computer science vs unemployment dynamics in economics).

Models of employee-vacancy matching often limit search scope based on the bipartite graph of workers and employers. In [3] people can only apply to firms employing their friends and firms are capable of hiring multiple workers. They focus on how quickly the matching process under these conditions converges to locally stable solutions and the total utility loss due to limited information. Another approach is to limit information over social network of individuals [4], studying a simulation where individuals are randomly unemployed, information about job openings is randomly dispersed throughout the social network, and unemployed individuals use information about openings themselves while the employed pass the information to friends. This study focuses on how initial network conditions, like having unemployed friends, can lead to the perpetuation of disadvantage within a marginalized group. Within networks search scope can also be limited based on distance across the network. While [5] and [3] consider matching within the 1-hop region of the network [6] focus on the 2-hop case, allowing random matching instead of ranked matching, and studying the time needed to find a locally stable match under these conditions.

An important question in matching problems with search scope limited by social networks is how to generate the network. This depends, in part, on the type of phenomenon one wishes to model. For example, rather than taking a *social* network approach, [7] look at an *exchange* network. In this model, when agents consider whether to form a new edge, they consider trade, and the optimal mix of two goods available for consumption. Since the model assumes ties are "costly," the network tends towards nodes linked in sparse linear chains of just a few nodes. This is extremely different from the studies using social networks, that tend to have many clusters and popular central nodes.

It might be the case that in marriage or job matching the counter-party's preferences are important. For example, if one individual is stably matched to someone who regards them as their "worst feasible match," the partner's dissatisfaction could lessen the individual's satisfaction. The same could hold true in a job, where workers want the best possible job they can get, but also want a satisfied employer. As such, some studies adjust not just search scope, but preference scope [5], allowing agents to incorporate others' preferences. [1]

[1] For a broad overview of matching limited by social networks see [8].

A second way to make matching algorithms more realistic is to limit time. A classic example in this area is the "secretary problem." In the secretary problem agents have no information about potential matches at the beginning of the process, and they gain information at the exact moment where they propose (or reject) a match, i.e. the employer has to decide whether to make an offer or not at the end of each interview [9]. One can also limit time and scope jointly. For example, [10] consider a situation firms first decide which potential employees to gather information about, and then use a standard matching algorithm with limited information, i.e. similar to the secretary problem, but the employer can interview a pool of people before making an offer.

There are other ways to limit information in matching, beyond time and social networks, that are not the focus of this paper. Some models allow information to be revealed over slow repeated interactions [11], some allow noisy signaling [12, 13], and others limit information based on past matching algorithm steps [14], i.e. excluding potential matches who had earlier been provisionally matched to someone to someone who had been provisionally matched to someone to whom they themselves had been provisionally matched.

Existing papers have not considered the case of jointly limiting search scope and time, which we do in this paper, measuring its negative effect on match quality and on time to convergence. Results suggest that jointly limiting scope and time does not make matches worse, as one might expect, but rather that limiting scope can partially compensate for losses due to limiting time. The model is designed to mirror dating in high school or college where people know those in their social networks and have preferences about who they would go out with. (This could also be considered analogous to a job search upon graduation or at the end of a fixed term job, where the individual knows about some subset of matching options and is seeking a match based on the information available within a short time frame.)

3 Method

3.1 Baseline Experiment: G-S Algorithm with Full Information

We simulate a pool of men and women randomly assigning four characteristics per agent: attractiveness (normal distribution, $\mu = 5$, $\sigma = 1$), intelligence (normal distribution, $\mu = 5, \sigma = 1$), preference for partner's attractiveness $\alpha =$(uniform distribution $(0, 1)$), and preference for partner's intelligence $(1 - \alpha)$. Men and women rank all members of the opposite sex by calculating their potential match utility using the “Cobb Douglas” utility function in equation 1. The importance of intelligence and attractiveness were constrained to sum to 1, meaning there are “constant returns to scale,” or doubling the partner’s attractiveness and intelligence would double one’s happiness. Following ranking, a G-S algorithm is implemented; men make offers to their most preferred women they have not yet proposed to, while women accept offers if they or single or if the offer is preferable to their current partner.

$$u_j = a_i^{\alpha_j} * s_i^{1-\alpha_j} \tag{1}$$

u_j	Utility of agent j
a_i	Attractiveness of potential partner i
α_j	Importance of attractiveness in partner for j
s_i	Intelligence of potential partner i
$1 - \alpha_j$	Importance of intelligence in partner for j

3.2 Limited scope with social networks

The most realistic context in which one could collect empirical data simultaneously on social networks and dating histories is within a limited school environment, where social boundaries are starkly delineated. This simulation reflects such an environment with the intent that simulation results could be compared with other empirical studies.

Social networks are characterized by a few standard characteristics. First, there is “homophile,” meaning that people tend to have friends who are like them. Second, social networks are characterized by clustering and triads, meaning that people form cliques. Third, social networks have skewed edge distributions, meaning that there are a few very popular people.

There are currently two main approaches to predicting the probability of edge formation in a social network: the “exponential graph model” (ERGM) and the “actor based model” [15, 16]. The two models are very different in their basic design, with the ERGM predicting the presence of edges at one point in time, while the actor based model is an agent-based model that moves in incremental time steps, allowing edge formation to interact with behavioral change (the same type of simulation as in this paper). Both can use equation 2 to model the probability of an edge, where the x_i variables indicate individual characteristics, x_{ij} indicates dyadic variables, D_i indicates an individual's current degree, and T_{ij} indicates triads, or mutual friends. The estimated coefficients, $/beta\_$, indicate the strength of each effect.

$$f = \beta_0 x_i + \beta_1 x_{ij} + \beta_D D_i + \beta_T T_{ij} \qquad Pr(i-j) = \frac{e^f}{1+e^f} \tag{2}$$

Ideally models are fit using multiple networks that can be considered copies of one another. As such, our simulated network is loosely based on estimations by [17], which use Adolescent Health Data predicting the presence of edges in 59 American High Schools where the median school was about 1000 people, about 300 per grade. Their model considers the relative effects of homophily based on race and age, and triad closure. Since the data is cross sectional popularity effects are not included. The pared down simulation used here cannot strictly match this data, but the general outline is comparable. As such, the simulation generates 150 men and 150 women with an average degree of 4, so as to be comparable to one school grade. Since degree distributions were skewed in the original data, we also added an additional popularity (preferential attachment) mechanism. Since characteristics like race and triad closure had about the same size effect, here we effect to be about equal too. While it is clear that the coefficients for homophily and edge closure should be about equal, the absolute level is unclear. We created 2 individual characteristics (we will call them attractive and intelligent) and then tuned coefficient by considering three ideal cases: 2 average people, 2 people with the maximum characteristics predicting a friendship edge (both attractive, smart, with 4 friends each and three friends in common) and a third case of 2 people unlikely to be friends (opposite extreme characteristics, no friends, and no friends in common). Using coefficients of .03 for attractive & smart, -.1 for dyadic attractive & smart differences, .05 for preferential attachment, and .15 for triad closure, the three dyads have probabilities of .49, and .87, and .24 respectively. The network is constructed by randomly picking two people, calculating the probability, p, of friendship, drawing a random number, n, from uniform(0,1), adding an edge if $p > n$, and repeating until the average degree in the simulated network is less than four. Once the social network is constructed, the men and women calculate their rankings, this time with men only making offers to women within x jumps of their social network with only y chances to make an offer.

4 Results

Figure 1 shows the utility of men and women by experimental condition, plotted over the number of offers men could make. The two steepest lines (coupled black and grey solid), shows men and women's utility in the full scope search. We can see that under this condition, utility for both men and women starts low, but increases steadily. At about 10 offers the full scope search converges to the stable solution. As we would expect, given that the Gale-Shapely algorithm is male-optimal, the male line (black) is parallel and consistently above the female (grey) line, indicating that men always do better under the full scope search. Under the 3-hop search condition (the paired blue and pink dash-dot lines), we see that utilities start somewhat higher than in the full-scope search, but that the stable solution is found more quickly (at about 7 offers) with utilities somewhat lower. In the 3-hop search the male advantage persists for up to about 5 offers, after which point it disappears. In the 2-hop search match utilities start even higher, but the stable solution is even sooner (at about 3 offers), with utilities even lower and almost no male advantage. By the time we get to the 1-hop search, the locally stable solution is found after just one round of offers with extremely low match utilities and no male advantage. In sum, the more time is limited, the smaller scope search is better; with one round 1-hop search dominates, between 1 and 4 rounds 2-hop dominates, from 4-9 rounds 3-hop dominates, after which a full-scope search dominates. What is particularly interesting about this result is that the assigned variables that impact *both* match quality and social network are intelligence and attractiveness, while the importance of one's partner's intelligence and attractiveness is only relevant for match quality. As such, there is no reason to expect that limiting scope should compensate for efficiency losses due to limiting time, but this is the case. Taking an anthropological perspective, these results might suggest that in cultures where people are expected to marry young, meeting a partner through close social networks makes sense, but as search time increases, it makes more sense to look more broadly.

Which characteristics predict satisfaction with one's match within each experiment, and how do those effects vary across experiments? Results from OLS regres-

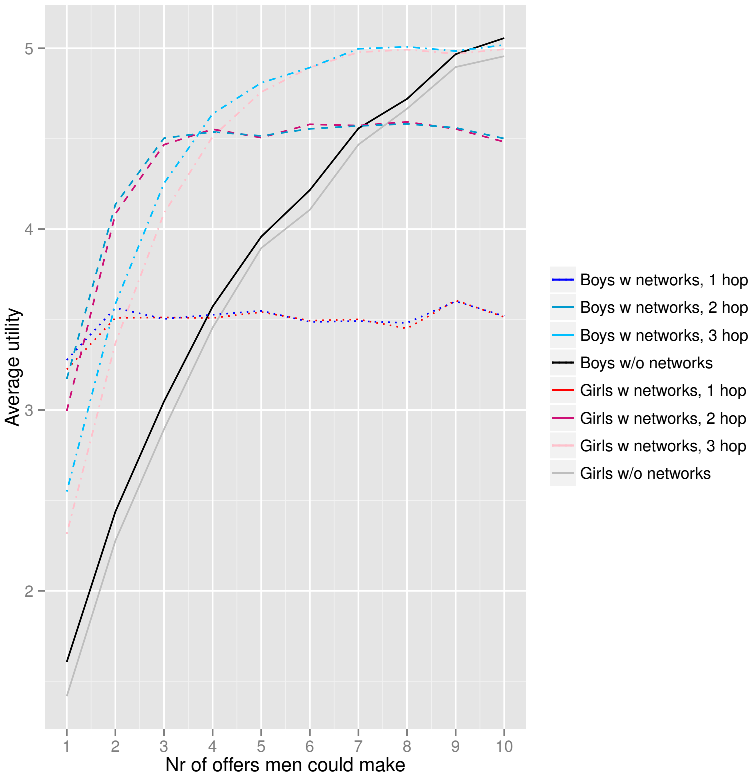


Fig. 1. Utility for men and women by experiment, over offer opportunities

sions on utility are illustrated in table 1 and for purpose of comparison, the coefficients are plotted in figure 2. The results show, as expected from the descriptive results in figure 1, that the female disadvantage inherent in the Gale Shapely algorithm are reduced in more constrained searches. We also expect that with a broader search scope, those who are more attractive would have the chance to benefit from their personal traits, which is the case. We would also expect that the individual degree matters more with more limited searches, since a broad personal network could compensate for limited search scope, which is also the case. Most interestingly, we find that personal traits matter more for women, and that these personal traits matter more with broader scope search, i.e. for smart attractive women, a broader search scope pays off relatively more. This effect breaks down such that unattractive (unintelligent) men do better than equally unattractive (unintelligent) women, while attractive (intelligent) women do better than equally attractive men.

Regressions were also run examining full interactions between the number of offers and individual characteristics, between number of offers and degree, number of offers and hops, and degree and hops, as well as with squared terms for friendship and hops. While results were all significant at the .001 level (simply because so many simulations were run), the size of the effects was very small, and so results are not illustrated. Results suggested that in a one-hop search those with a higher degree have an advantage, but this can be compensated with broader searches; optimum match utilities are reached for those in a 2-hop search with 7 or more friends. By the 3-hop search, a higher friendship degree ceases to help in finding a match. While those who are more attractive or smart always find better partners, in 2 and 3-hop searches, the benefit declines. Low quality men also benefit from limited search scope and time. High quality women benefit from longer search times.

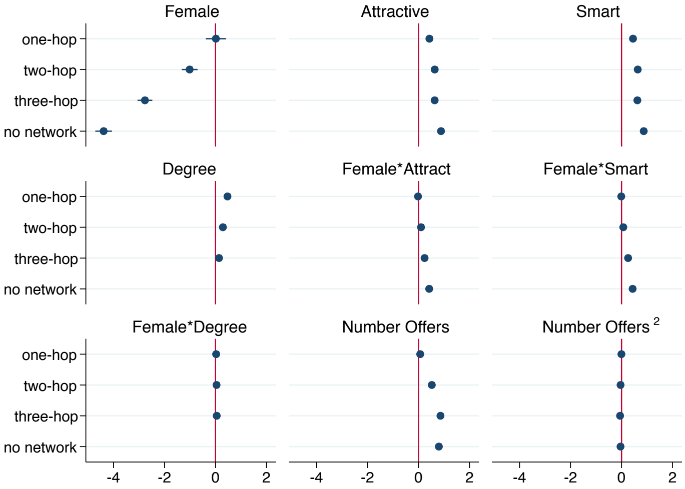


Fig. 2. Predictors of utility by characteristics, within experiment (OLS)

In sum, results are as follows:

- Limited search scope partially compensates for limited time.
- Male advantage declines with constrained search scope but not with limited time.
- Higher quality individuals have higher quality matches.
- Higher quality individuals benefit with broader scope, not with more time.
- A high friendship degree can compensate for limited scope.
- A high friendship degree is more of an advantage with more time.

	(1-hop)	(2-hop)	(3-hop)	(no) network
Female	0.0153	-1.014***	-2.768***	-4.387***
	(0.08)	(-6.41)	(-18.70)	(-26.29)
Attractive	0.428***	0.636***	0.634***	0.883***
	(21.95)	(41.75)	(44.08)	(53.90)
Smart	0.450***	0.636***	0.622***	0.870***
	(22.94)	(41.63)	(43.48)	(52.53)
Degree	0.470***	0.290***	0.138***	
	(48.81)	(38.71)	(19.85)	
Female∗Attractive	-0.0191	0.0998***	0.239***	0.420***
	(-0.69)	(4.62)	(11.78)	(18.02)
Female∗Smart	-0.0105	0.0692**	0.258***	0.436***
	(-0.38)	(3.23)	(12.81)	(18.69)
Female∗Degree	0.0249	0.0408***	0.0484***	
	(1.84)	(3.87)	(4.92)	
Number Offers	0.0680**	0.522***	0.862***	0.799***
	(3.16)	(31.34)	(54.84)	(44.09)
Number Offers^2	-0.00463*	-0.0381***	-0.0570***	-0.0394***
	(-2.42)	(-25.80)	(-40.90)	(-24.55)
Constant	-2.959***	-4.576***	-4.904***	-7.816***
	(-19.47)	(-38.54)	(-44.11)	(-62.59)
N	25000	25000	25000	25000
R^2	.2243	.3450	.4328	.5199

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1. OLS regression predicting match utility within experiment

5 Importance and Future Research Questions

The main finding is that a social network with orthogonal edge-generation criterion, can create better quality matches given limited time. The puzzle is why limiting the matching pool increases utilities with limited match time, when the criterion for generating social network edges is orthogonal to the match utility. Further research is needed to understand this completely, and to consider the range of imposed networks that would generate this phenomenon.

This matching experiment was run on a relatively small network of only 150 men and 150 women, as to be comparable with data collected in a school environment. As such, a stable match with full information was found in just 10 steps. In a larger network, the processes found here might be constrained or dampened.

References

1. Gale, D., Shapely, L.: College admissions and the stability of marriage. The American Mathematical Monthly **69**(1) (1962) 9–15
2. Gunter, J.H., Hortacsu, A., Ariely, D.: What makes you click? mate preferences in online dating. Quantitative Marketing and Economics **8**(4) (2010) 393–427
3. Arcaute, E., Vassilvitskii, S.: Social networks and stable matchings in the job market. Technical Report 0910.0916, arXiv (2009)
4. Calvo-Armengol, Jackson, M.O.: The effects of social networks on employment and inequality. American Economic Review **94**(3) (2007) 426–54
5. Anshelevich, E., Bhardwaj, O., Hoefer, M.: Friendship and stable matching. Technical report, Proc. 21st European Symposium on Algorithms (ESA) (2013)
6. Hoefer, M., Wagner, L.: Locally stable marriage with strict preferences. Automata, Languages, and Programming, Lecture Notes in Computer Science **7966** (2013) 620–631
7. Jackson, M.O., Watts, A.: The evolution of social and economic networks. Journal of Economic Theory **106** (2002) 265–295
8. Stoovel, K., Fountain, C.: Matching. In Hedtröm, P., Bearman, P., eds.: Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford (2009)
9. Freeman, P.R.: The secretary problem and its extensions: A review. International Statistic Review **51**(2) (1983) 189–206
10. Lee, R., Schwartz, M.: Interviewing in two sided matching markets. Technical report, American Economic Association Annual Meeting (2008)
11. Das, S., Kamenica, E.: Two-sided bandits and the dating market. Technical report, Proceedings of the Nineteenth International Joint Conferences on Artificial Intelligence (2005)
12. Chade, H., Smith, L.: Simultaneous search. Econometrica **74**(5) (2006) 1293–1307

13. Hoppe, H.C., Moldovanu, B., Sela, A.: The theory of assortative matching based on costly signals. Review of Economic Studies **76**(1) (2009) 253–281
14. Bearman, P.S., Moody, J., Stovel, K.: Chains of affection: The structure of adolescent romantic and sexual networks. American Journal of Sociology **110**(1) (2004) 44–91
15. Snijders, T.A.: The statistical evaluation of social network dynamics. Sociological Methodology **31** (2001) 361–395
16. Snijders, T.A., Steglich, C.E.: Representing micro-macro linkages by actor-based dynamic network models. Sociological Methods & Research **00** (2013) 1–50
17. Goodreau, S.M., Kitts, J.A., Morris, M.: Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. Demography **46**(1) (2009) 103–125

Problem Complexity in Parallel Problem Solving

Sebastian Herrmann*, Jörn Grahl and Franz Rothlauf

Dept. of Information Systems and Business Administration,
Johannes Gutenberg-Universität,
Jakob Welder-Weg 9, 55128 Mainz, Germany
http://wi.bwl.uni-mainz.de
{s.herrmann,grahl,rothlauf}@uni-mainz.de

Abstract. Recent works examine the relationship between the communication structure and the performance of a group in a problem solving task. Some conclude that inefficient communication networks with long paths outperform efficient networks on the long run. Others find no influence of the network topology on group performance. We contribute to this discussion by examining the role of problem complexity. In particular, we study whether and how the complexity of the problem at hand moderates the influence of the communication network on group performance. Results obtained from multi-agent modelling suggest that problem complexity indeed has an influence. We observe an influence of the network only for problems of moderate difficulty. For easier or harder problems, the influence of network topology becomes weaker or irrelevant, which offers a possible explanation for inconsistencies in the literature.

Key words: Problem solving, networks, computational social science, group performance, multi-agent modeling.

1 Problem Solving in Groups

Humans routinely assemble into groups to solve complex problems. As they exchange ideas and approaches, the performance of the group in solving a problem depends on the individual performance of the group members and on the type and structure of the collaboration. A question that has widely been discussed in literature is how the communication network's structural properties influences group

* Corresponding author

performance. The topic is of interest to psychology, sociology as well as management and organization science. Some studies find that complete freedom of communication can be more limiting than restricted communication patterns [1–3]. Other studies come to conflicting conclusions or a more differentiated view where the optimal group structure depends on the tasks applied in the experiments [4–6]. Most studies draw conclusions from small group experiments or observational data. Recent research in the computational social sciences tries to overcome this limitation by running multi-agent models or web-based experiments with large groups [7].

Lazer & Friedman [8] try to identify superior network structures in a scenario called *parallel problem solving*. Here a set of roughly equivalent actors attempts to solve a complex problem. They conduct an agent-based simulation (hereafter referred to as *LF-model*). The agents search for good solutions in an NK-landscape [9], a well-established model for problem representation in organizational theory [10]. A solution is represented as a binary string of N bits and has a score that agents try to maximize. They can *explore* from a given solution to another by flipping a random bit. Alternatively, they can *exploit* solutions of their network neighbors by copying them if they are better. The problem space is multi-modal and moderately rugged. Search can lead to a state in which no further improvement is achievable even though that state is not the global optimum. Lazer & Friedman average the scores of all agents to measure the performance of the group. They compare the group performance for several networks. The networks are characterized by differences in efficiency (their average path length). Networks with higher efficiency (smaller path lengths) are able to disseminate information faster than less efficient networks. Their results suggest that efficiency is beneficial for short-run performance. On the long run however, less efficient networks perform better.

Mason & Watts [12] conduct a series of web-based experiments with human subjects (hereafter referred to as *MW-model*). The subjects play a networked game with the objective to select points for oil-drilling on a map. They do not know where the good and where the bad oil wells are positioned on the map, but they are able to see the coordinates and earnings of their network neighbors. Contrary to [8] they find that the average path length of the network that connects the subject is negatively correlated with the mean earnings achieved by the group. Put differently, networks with shorter paths perform better. They do not find a significant difference in the probability to find the best solution, even though efficient networks have a slightly higher performance.

Lazer & Friedman conclude that an efficient network negatively affects information diversity which in turn negatively affects group performance. Higher network efficiency leads to lower average scores. Mason & Watts find no significant difference between the networks in terms of success probability. As efficient networks have higher earnings on average, they presume that—in case of doubt—efficient information flow can only be advantageous. We believe that both models provide valuable insights into basic mechanisms in collaboration and indicate that communication structure is important. Nevertheless, it is an open question whether the results are substantially conflicting or if they are caused by different assumptions or experimental conditions.

We believe that a deeper understanding of the network influence on group performance requires that additional factors are included into the analysis. A glimpse into small groups studies reveals that several other factors have already been discussed. Heise & Miller [13] find that for simple problems, groups with more communication channels produce less errors, for intermediate problems more tightly organized groups are better. For high complexity, there is no difference at all between groups. Others find the opposite [14]: for simple problems, group structures makes no difference in accuracy, but for complex problems, less tightly structured groups produce less errors. Even though these studies have contradictory findings, both authors agree that there is an influence of the task complexity. Also Lazer & Friedman propose that complexity should be taken into consideration in their model. Mason [15] suspects that task complexity ("the potential for local maxima") could be a necessary condition for the superiority of networks with long paths in the long run. Hence, the goal of this paper is to provide first insights into how problem complexity moderates the network influence on group performance.

2 Experiments and Results

In the original experiments on the LF-model the agents have to solve problems of moderate complexity. We conduct a series of experiments with a modified LF-model. Our intention is to test the impact of network structure on problem-solving capability for different levels of task complexity. A side condition is to make as little changes as possible to the original model, thus the group size of 100 agents per network, behavioral parameters etc. are left unchanged. Yet a presumption we have to reconsider is that of the operationalization for group performance. Comparing the original model to real-world challenges—e.g. research on composition of an effective pharmaceutical—we see a lack of construct validity: these problems are analogous to combinatorial optimization problems, where the detection of the optimal solution is desirable (so-called conjunctive tasks) [16]. Under this assumption, the average degree of target achievement of the group members (as assumed in [8]) is rather inaccurate. Hence, we define group performance as the probability to find the optimal solution.

A rigorous approach to test the impact of network structure for a particular task should consider all networks that can be evolved for a given number of agents. This is impracticable because the number of possible networks grows exponentially with the number of actors. We therefore measure network influence as the difference in long-term performance between a linear and a fully connected network. These structures have been studied in [8] and they represent both extremes of the concept of average path length. A linear network has the highest average path lengths attainable with a given number of nodes, a totally connected one has the lowest average path lengths.

A benefit of NK-landscapes is that their complexity can be tuned by the parameter K, which determines the number of interdependencies between the N binary decision variables. A value of $K = 0$ results in a smooth uni-modal, easy-to-solve space, whereas $K = N - 1$ causes a maximally complex space. We test instances of the LF-model for a variety of difficulty levels. For this purpose 100 random NK-landscapes are generated for each $K \in [0, 19]$ ($N = 20$ bits). We execute 100 repetitions for each space and then count in how many cases at least one agent finds the optimal solution.

Fig. 1 shows the result. For intermediate problem difficulty, the linear network outperforms the totally connected network on the long run, but is inferior on the short run. This is in accordance with the original model, even though our metric for group performance is different. However, the more we increase or decrease problem complexity from this point, the differences in performance between both networks vanish: the network influence falls towards zero as K approaches 0 or 19. We observe a clear influence of network structure on the groups' success probability for intermediate problem difficulty only and reduced or even absent influence

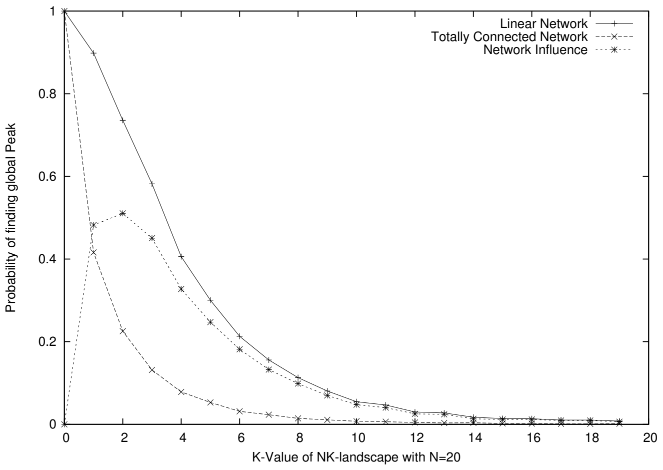


Fig. 1. The vertical axis shows the probability of finding the optimal solution for linear and totally connected networks in 100 random NK-landscapes. The horizontal axis shows the parameter K as a measurement of problem complexity, from very easy ($K = 0$) to very hard ($K = 19$). Network influence is the difference between the success probabilities of the two networks. Lines between points are for illustration purposes only.

to both extreme sides of the complexity metric. Imagining a line between the data points for network influence in fig. 1, a curved shape appears.

To explore if the curvilinear relationship is a peculiarity of the NK-landscape we let the agents solve a Traveling Salesman Problem (TSP). The goal of a TSP is to connect places on a map in a single route so that the total distance for the entire tour is minimized. Using the TSP is a first step into the direction of understanding network influence on more realistic settings. Full implementation details are in [18]. In our case the complexity of the TSP is constant for a given number of cities [19]. Hence, to vary the complexity we vary the number of cities from 1 to 20. The success of the group is again the probability that at least one member finds the optimal solution. The results confirm our previous finding (fig. 2). We observe differences in performance only for intermediate problem complexity and diminishing effects for extreme difficulty levels. This suggests that the curvilinear relationship between problem complexity and network influence is not a peculiarity of the NK-landscape.

3 Summary and Conclusion

Our analyses suggest that the influence of network structure on group performance is more nuanced than previously thought. There is high evidence for a curvilinear moderating influence of task difficulty on the relationship between a

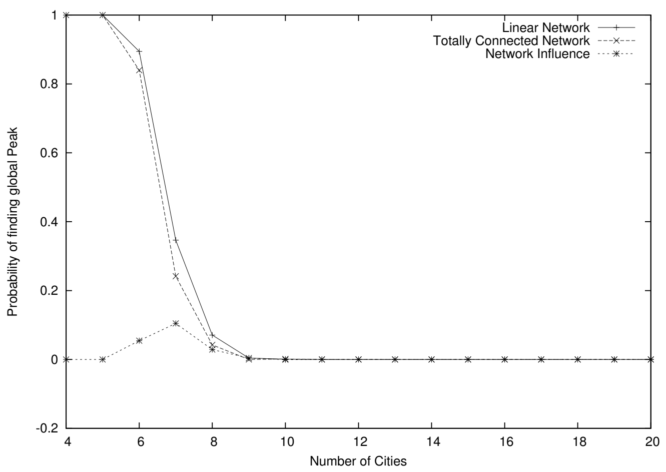


Fig. 2. The vertical axis shows the probability of finding the optimal solution for linear and totally connected networks in 100 random TSP-instances. The horizontal axis represents the number of cities as the parameter for problem complexity. Network influence is defined as difference between the success probabilities of the two networks. Lines between points are for illustration purposes only.

group's network structure and its performance. People may be able to solve very easy problems, independent from their structure of collaboration. For very hard problems, collaboration in any form may have no effect either as the problem is so unstructured that the exploitation of promising ideas from others will not lead actors towards a direction in which the optimal solution can be found. For intermediate problems, the network structure affects a groups ability to find appropriate solutions and an adequate balance of exploration and exploitation is necessary. Our findings also indicate that recent results on the influence of network structure may not be as contradictory as they appear. Insignificant network effects could be an artifact of the task complexity. For instance, Mason & Watts [12] found no significant difference between networks in terms of success probability. It could be possible that the problem used in their setup is too easy or too hard for a network influence to play a significant role.

It is difficult to draw managerial conclusions from computational studies. Future research should try to replicate our findings with human subjects.

References

1. Leavitt, H.J.: Some Effects of Certain Communication Patterns on Group Performance. J. Abnorm. Soc. Psych. p. 46, 38 (1951)

2. Guetzkow, H., Simon, H.A.: The Impact of Certain Communication Nets upon Organization and Performance in Task-oriented Groups. Manage. Sci. 1 (3–4), pp. 233–250 (1955)
3. Cohen, A.M., Bennis, W.G., Wolkon, G.H.: The Effects of Changes in Communication Networks on the Behaviors of Problem-solving Groups. Sociometry, 25, p. 177 (1962)
4. Shaw, M.E.: Some Effects of Unequal Distribution of Information upon Group Performance in Various Communication Nets. J. Abnorm. Soc. Psych. 49, pp. 547–553 (1954)
5. Mulder, M.: Communication Structure, Decision Structure and Group Performance. Sociometry 23, pp. 1–14 (1960)
6. Carzo, R.: Some Effects of Organization Structure on Group Effectiveness. Administrative Science Quarterly 7, pp. 393–424 (1963)
7. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Devon Brewer, Christakis, N., Contractor, N., Fowler, J., Myron Gutmann, Jabara, T., King, G., Macy, M., Roy, D., van Alstyne, M.: Computational Social Science. Science 323, pp. 721–723 (2009)
8. Lazer, D., Friedman, A.: The Network Structure of Exploration and Exploitation. Admin. Sci. Quart. 52, pp. 667–694 (2007)
9. Kauffman, S., Weinberger, E.: The NK Model of rugged fitness landscapes and its application to the maturation of the immune response. J. Theor. Biol. 141 (2), pp. 211–245 (1989)
10. Levinthal, D. A.: Adaptation on Rugged Landscapes. Manage. Sci. 43 (7), pp. 934–950 (1997)
11. Ryan, C., Keijzer, M., Ochoa, G., Tomassini, M., Vérel, S., Darabos, C.: A Study of NK Landscapes' Basins and Local Optima Networks. In: GECCO Proceed. (2008)
12. Mason, W., Watts, D.J.: Collaborative Learning in Networks. Proc. Natl. Acad. Sci. 109, pp. 764–769 (2012)
13. Heise, G.A., Miller, G.A.: Problem Solving by Small Groups using Various Communication Nets. J. Abnorm. Soc. Psych. 46, pp. 327–335 (1951)
14. Shaw, M.E.: Some Effects of Problem Complexity upon Problem Solution Efficiency in Different Communication Nets. J. Exp. Psych. 48, p. 211 (1954)
15. Mason, W.: Human Compution as Collective Search. In Michelucci, P., ed.: Handbook of Human Computation. Springer, New York (2013)
16. Steiner, I.D.: Group Process and Productivity. Social Psychology. Academic Press, New York (1972)
17. Kallel, L., Naudts, B., Reeves, C.: Properties of Fitness Functions and Search Landscapes. In Kallel, L., Naudts, B., Rogers, A., eds.: Theoretical Aspects of Evolutionary Computing. pp. 175–206. Springer Berlin Heidelberg (2001)
18. Lazer, D., Gomez, C.: Global and Local Diversity and Systemic Network Performance. Ann. Meet. ASA (2012)
19. Stadler, P.F.: Landscapes and their correlation functions In: J. Math. Chem. 20(1), pp. 1–45 (1996)

Reviewed Papers

FGENET 2014

A Perspective on the Future Retail Energy Market

Michael Höfling, Florian Heimgärtner, Benjamin Litfinski and Michael Menth

University of Tübingen, Chair of Communication Networks,
Sand 13, 72076 Tübingen, Germany
{hoefling,florian.heimgaertner,menth}@uni-tuebingen.de,
benjamin.litfinski@student.uni-tuebingen.de

Abstract. Electrical energy will be more expensive and less predictable in the near future. A leading factor in this trend is the mass deployment of renewable energy sources. In this paper, we sketch the structure of the electrical energy grid and explain why power supply will be more demanding in the future. More volatile energy prices and small energy suppliers will create more activity on the retail energy market (REM). We present a perspective on the future REM that calls for communication support to satisfy the information needs of the market participants.

Key words: Smart grid, retail energy market, market structure

1 Introduction

Power generation is currently changing from a centralized system with predictable and controllable outputs to a system integrating distributed energy resources (DERs) including weather-dependent renewables. Such renewable energy sources are hard to predict and impossible to control [1,2]. There is strong societal pressure to protect the environment, explore cleaner alternatives to fossil fuels, improve energy efficiency, and reduce carbon emissions. The downside is that we will face variations in supply, with periods of higher or lower renewable energy offers. The deficit must be compensated by other, more expensive energy sources to avoid outages. This will affect future markets for electrical energy.

In other words, future prices for electrical energy will fluctuate more than today. Nevertheless, a normal household will still be able to buy electrical energy for a fixed price per period from a retailer, but at increased cost. Consumers may be better off buying power directly from prosumers or DERs than from retailers on the retail energy market (REM), thus taking advantage of lower prices at certain times, and possibly shifting parts of their demand to other times of day, which is a desired behavior [3].

Today, DERs like photo-voltaic (PV) panels or wind farms sell their generated power for a fixed, subsidized price. When the fixed-price contract model expires, they may sell their energy on the REM, too. As a consequence, the future REM for electrical energy will have many more participants and see more volatile prices than today, creating the need for new trading infrastructures [4,5,6].

2 Structure of Power Grids

Power grids are hierarchically structured. They can be divided broadly into three different domains: power generation, power transmission, and power consumption. In addition to the domains, there are four different voltage levels: extra high-voltage (EHV), high-voltage (HV), medium-voltage (MV), and low-voltage (LV). Substations transform between the voltage levels. Figure 1 shows the structure of a typical power grid.

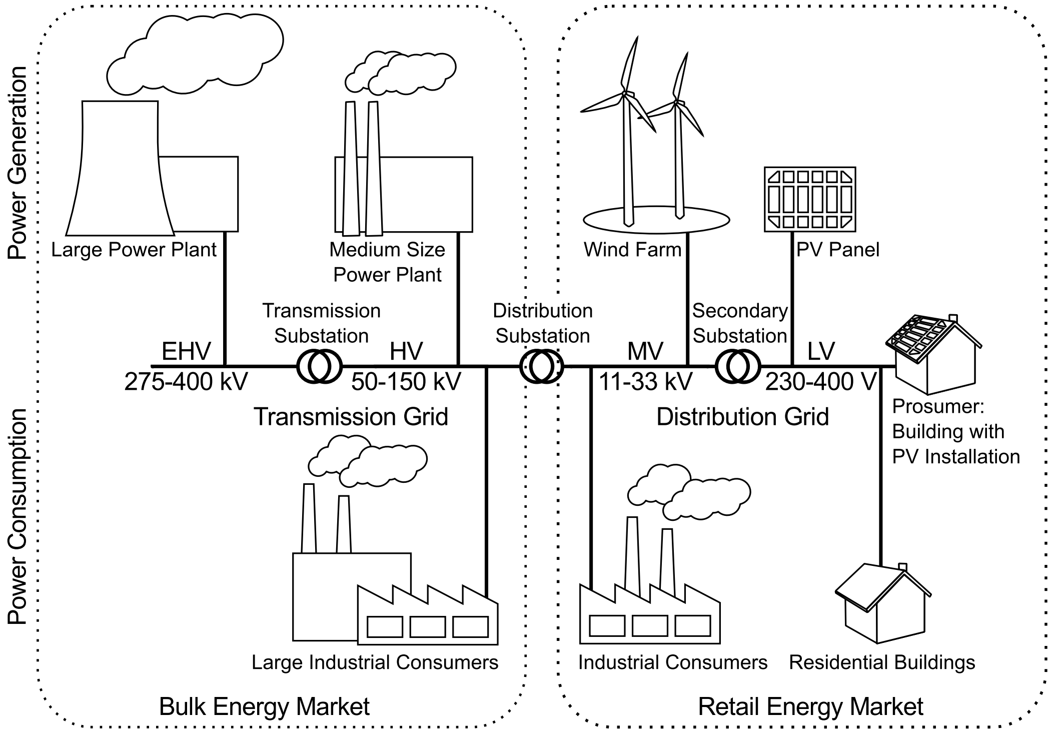


Fig. 1. The structure of a typical power grid including the general boundaries of the electrical energy market. Power is generated at the top, transfered over the transmission and distribution grid, and consumed at the bottom. Prosumers are positioned in-between the generation and consumption domain as they are part of both domains. Voltage levels decrease from left to right, i.e., from EHV level to LV level.

The *power generation* domain consists of power plants, e.g., coal, nuclear, or hydro-electric plants, but also DERs, e.g., wind farms or PV panels. The transmission grid transports power over long distances, sometimes even across international borders. The distribution grid facilitates regional distribution of power. Combined, both grids form the *power transmission* domain. The *power consumption* domain covers all service locations consuming power, e.g., industrial consumers and residential buildings. Prosumers are special entities since they belong to both the power generation and the power consumption domain. They may produce power and feed-in to the grid, but they may also consume power. The normal power flow is unidirectional: top-down from the generation domain to the consumption domain, and from left to right in the transmission domain, i.e., from EHV level to LV level. With the increasing number of DERs, bidirectional power flow inside the transmission domain is possible, e.g., from LV to MV level.

3 Today's Bulk and Retail Energy Market

We now take a closer look at today's electrical energy market and its market mechanisms. From an economic point of view, electrical energy is a commodity which can be bought, sold, and traded. Depending on which participants interact with each other on what voltage level, we differentiate between two markets: bulk energy market (BEM) and REM. Figure 1 illustrates the boundaries of BEM and REM. In practice, there is no sharp border between both markets.

The *BEM*, sometimes referred to as wholesale market, consists of three major participants on the EHV, HV, and MV level of the power grid: suppliers of energy, retailers, and large consumers. Competing suppliers of energy offer their electrical energy on the BEM to retailers or large consumers of electrical energy, e.g., aluminum plants. Large consumers buy electrical energy through the BEM directly. Energy trading normally takes place on trading platforms similar to the stock exchange. However, BEM transactions are also possible without involving a trading platform. An example for a BEM trading platform is the European Energy Exchange (EEX) [7], which spans Germany, France, Austria, and Switzerland. Typical time scales for BEM transactions on the EEX vary between hours and years.

The *REM* consists of two major participants on the MV and LV level of the power grid: retailers and clients. *Retailers* buy electrical energy on the BEM, and resell it through the REM to clients not participating in the BEM. *Clients* buy or sell electrical energy on the REM. Examples for clients are consumers, prosumers, and DERs. The REM enables clients to choose their electrical energy supplier from competing retailers. In contrast to the BEM, energy on the REM is not traded directly between all participants but indirectly through the retailer. That is, clients can buy or sell energy only through retailers.

All transactions between consumers of energy and suppliers of energy on the REM are called retail energy transactions (RETs). Today's RETs include three consecutively executed phases: retailer selection by clients, delivery of electrical energy, and accounting for the delivered electrical energy. While the meaning of each phase is self-explanatory their exact realization in today's REMs is subject to country-specific legislations. Today's RETs are based on fixed-price contract models, i.e., a client buys or sells a certain amount of electrical energy at a fixed price per energy unit for a specified period on the REM. The time scale of today's RETs is given by the accounting period of the electricity contract, e.g., one month, one year, or even longer. However, no generally agreed fixed time scale for today's RETs is given in the literature.

4 Future Retail Energy Market

In the future REM, any participant will be able to trade energy. Instead of a fixed-price contract model, consumers will have dynamic pricing based on predicted supply and demand [3]. Electricity trading intervals will be on the order of minutes or hours, i.e., significantly shorter than today's accounting intervals [6]. As a consequence, the future REM will have many more participants and see more volatile prices than today. New trading infrastructures are necessary as enabling technology [4,5,6]. In the following, we briefly introduce the structure of the future REM, mention the concept of coalitions, and eventually sketch future RETs.

4.1 Market Structures

In the literature there exist various definitions of future market participants and their functions [2,4,6,8,9,10]. We provide a unified view thereof in Figure 2. The figure shows what a future REM may look like compared to today's REM. Besides additional participants, cash and energy flows, and communication flows will change. The future REM can be divided into five classes of participants: clients, aggregators (AGGRs), energy supply managers (ESMs), distribution system operators (DSOs), and regulators (not shown in the figure).

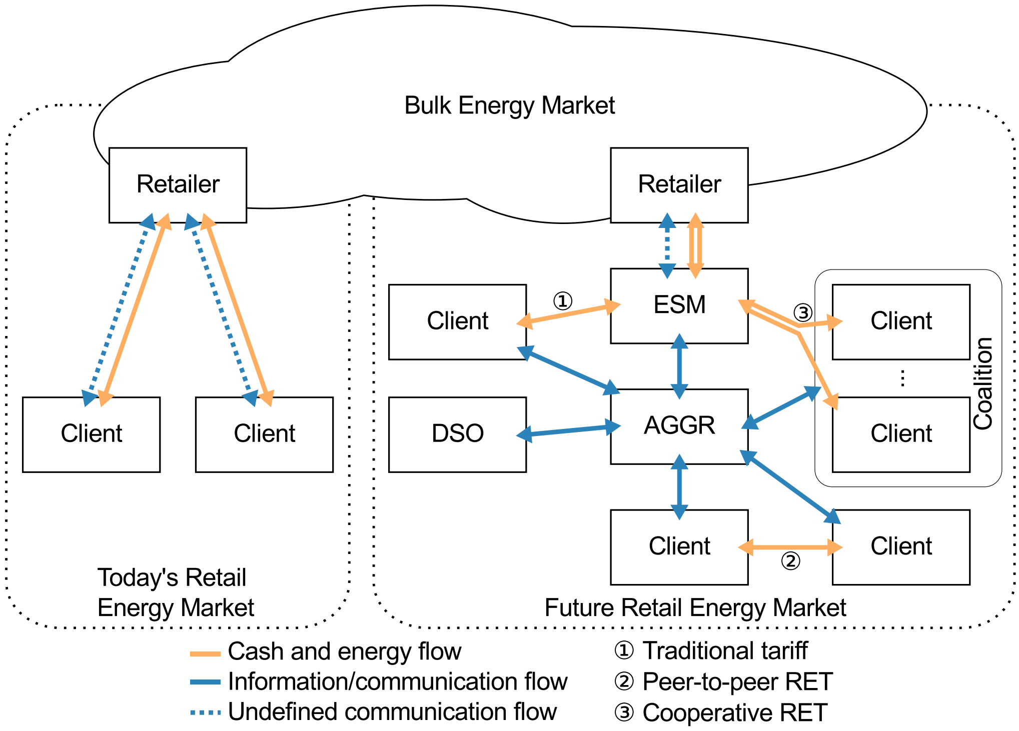


Fig. 2. Cash, energy, and communication flows in today's and the future REM. In today's REM, clients can only sell or buy energy through retailers. Trading between clients is only possible indirectly using retailers. In the future REM, in addition to traditional tariffs (1), AGGRs enable clients to directly trade their energy with each other (2). Groups of clients may form coalitions and participate in collaborative RETs (3), e.g., to maximize profit. ESMs guarantee energy balance inside distribution grids, while DSOs verify physical constraints of RETs.

Clients in the future REM cannot only buy and sell electrical energy from or to retailers, but they can also trade their electrical energy directly on the REM. They have to provide proper forecasts of their energy demand and supply, possibly based on weather forecasts if their power production is weather-dependent.

AGGRs supervise demand supply matching (DSM). They mediate between clients for DSM inside the distribution grid, and between clients and ESMs for DSM between the distribution grid and the transmission grid. AGGRs are the only authoritative entity in the future REM to initialize and supervise auctions, and they prevent trades that cannot meet physical constraints.

ESMs are responsible for balancing the energy in the distribution grid. For example, if the energy demands of distribution grids exceed their internal production, ESMs acquire additional electrical energy on the BEM to ensure proper energy supply in the distribution grids.

DSOs are control instances of distribution grids. They operate distribution grids and validate the outcomes of auctions, so-called power transaction plans. That is, if the outcome of an auction would lead to an unstable grid configuration violating physical constraints, the auction is invalidated and AGGRs may be asked to restart the auctions.

Regulators are independent authorities that determine or approve the electricity market rules, and monitor RETs to ensure compliance with regulations and rules.

4.2 Coalitions

Normally, each client acts as an individual participant in a RET. The minimum achievable profit by a single client is given by the so-called *self-value* [11]. The self-value depends on client-specific parameters, e.g., estimated weather-dependent energy production, or the geographical location of the client. The future REM introduces client coalitions to maximize client profits [11,12,13,14] or to create efficient virtual power plants [15]. Client coalitions are temporary groups of clients, not necessarily geographically close to each other, pursuing short-term common economic interests. Coalition formation is a distributed process which enables clients to find and agree on potential coalition partners. During coalition formation, each client calculates its self-value and disseminates it to all other clients through the AGGR. Coalition decisions are then made based on the self-values, i.e., each client independently determines whether a coalition with one or more clients matches its economic objectives.

From the market's perspective, coalitions are virtual clients with their own self-value participating in RETs. A virtual power plant is an example for such a client coalition, i.e., prosumers and DERs are aggregated into a virtual equivalent of a large power plant. Coalitions are included here because they are an active research area, but RETs are possible without coalitions as well, i.e., coalitions are an optional feature. We will use the term clients interchangeably for both clients and coalitions.

4.3 Future Retail Energy Transactions

The future REM supports three different types of future RETs: traditional tariff, peer-to-peer (P2P), and collaborative. Traditional tariff RETs are comparable with today's RETs based on fixed-price contracts. However, communication flows for traditional tariffs differ as shown in Figure 2. Clients communicate with retailers through AGGRs and ESMs. P2P RETs [4,16] are direct transactions between two clients which have been coordinated using the AGGR. Collaborative RETs [12,13,14,15] are transactions between coalitions and clients, or coalitions and coalitions.

In contrast to today's RETs, the retailer selection phase is replaced by a two-stage process consisting of *coalition formation* and *auctions* in future RETs. Coalition formation is optional as described in Section 4.2. The auction phase between clients is initialized and coordinated by the AGGR. That is, each client sends its demand and supply prediction to the AGGR which then matches the received demands and supplies. The outcome of the auction is a *power transaction plan* which is sent to the DSO for approval considering the physical constraints of the distribution grid. If the approval is successful, the AGGR sends a binding agreement to the clients. After the delivery of electrical energy, the accounting phase matches actual demands and supplies with their originally predicted values. Clients which did not fulfill their demand or supply prediction are penalized.

5 Conclusions

The evolution of power grids to smart grids leads to new technical, political, and economical challenges. In this paper, we presented a perspective on the future REM based on an extensive literature study. In Germany, projects like SESAM, DINAR and BEMI [5,8] already address energy control, management and trading. To enable the future REM, new trading infrastructures are needed. The FP7 project C-DAX [17] works on an information architecture for which the future REM is a use case. Further investigations of existing and future problems need interdisciplinary efforts of electrical engineers, computer scientists, and economists.

Acknowledgments The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7-ICT-2011-8 under grant agreement n° 318708 (C-DAX). The authors alone are responsible for the content of this paper.

The authors thank C. Mills for valuable input and stimulating discussions.

References

1. Borggrefe, F., Nüßler, A.: Auswirkungen fluktuierender Windverstromung auf Strommärkte und Übertragungsnetze. uwf UmweltWirtschaftsForum **17**(4) (2009)
2. Franke, M., Rolli, D., Kamper, A., Dietrich, A., Geyer-Schulz, A., Lockemann, P., Schmeck, H., Weinhardt, C.: Impacts of Distributed Generation from Virtual Power Plants. In: International Sustainable Development Research Conference. Volume 11., Helsinki, Finland (June 2005)
3. Eßer, A., Franke, M., Kamper, A., Möst, D.: Future Power Markets – Impacts of Consumer Response and Dynamic Retail Prices on Electricity Markets. WIRTSCHAFTSINFORMATIK **49** (2007)
4. Capodieci, N., Pagani, G.A., Cabri, G., Aiello, M.: Smart Meter Aware Domestic Energy Trading Agents. In: E-Energy Market Challenge Workshop. (2011)
5. Bendel, C., Nestle, D., Ringelstein, J.: Bidirektionales Energiemanagement im Niederspannungsnetz: Strategie, Umsetzung und Anwendungen. e&i Elektrotechnik und Informationstechnik **125** (2008)
6. Block, C., Collins, J., Ketter, W.: Agent-based Competitive Simulation: Exploring Future Retail Energymarkets. In: International Conference on Electronic Commerce, Honolulu, HI, USA (2010)
7. European Energy Exchange AG: European Energy Exchange (EEX) (2013) `http://www.eex.com/` (last visited November 2013).
8. Bendeli, C., Nestle, D., Ringelstein, J., Eßer, A., Möst, D., Rentz, O., Franke, M., Geyer-Schulz, A.: Marktmodell für ein dezentral organisiertes Energiemanagement

im elektrischen Verteilnetz - Grundlage für ein internetbasiertes Managementsystem. In: ETG-Kongress, Fachtagung Webbasierte Automatisierung in der elektrischen Energietechnik, Karlsruhe, Germany (2007)

9. Gkatzikis, L., Koutsopoulos, I., Salonidis, T.: The Role of Aggregators in Smart Grid Demand Response Markets. IEEE Journal on Selected Areas in Communications **31**(7) (July 2013)
10. C-DAX Consortium: C-DAX Deliverable D2.1: C-DAX Requirements - Use Case Descriptions for Domains 1, 2 and 3 and Derived C-DAX Requirements (April 2013)
11. Yeung, C., Poon, A., Wu, F.: Game Theoretical Multi-Agent Modelling of Coalition Formation for Multilateral Trades. IEEE Transactions on Power Systems **14**(3) (1999)
12. Contreras, J., Candiles, O., de la Fuente, J., Gomez, T.: Auction Design in Day-ahead Electricity Markets. IEEE Transactions on Power Systems **16**(1) (2001)
13. Hazard, C.J., Wurman, P.R.: The Game of Scale: Decision Making with Economies of Scale. In: International Conference on Electronic Commerce, Minneapolis, MN, USA (2007)
14. Corchero, C., Mijangos, E., Heredia, F.J.: A New Optimal Electricity Market Bid Model Solved Through Perspective Cuts. TOP **21**(1) (2013)
15. Chalkiadakis, G., Robu, V., Kota, R., Rogers, A., Jennings, N.R.: Cooperatives of Distributed Energy Resources for Efficient Virtual Power Plants. In: International Conference on Autonomous Agents and Multiagent Systems, Taipei, Taiwan (May 2011)
16. Capodieci, N.: P2P Energy Exchange Agent Platform Featuring a Game Theory Related Learning Negotiation Algorithm. Master's thesis, Universita degli Studi di Modena e Reggio Emilia (2011)
17. C-DAX Consortium: Cyber-secure Data And Control Cloud for Power Grids (C-DAX) (2013) `http://www.cdax.eu/` (last visited November 2013).

A Coupled Optimization and Simulation Model for the Energy Transition in Bavaria

Marco Pruckner[1], Christoph Thurner[2],
Alexander Martin[2] and Reinhard German[1]

[1] Computer Networks and Communication Systems, University of Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany,
marco.pruckner@cs.fau.de, german@cs.fau.de
[2] Economics, Discrete Optimization, Mathematics (EDOM), University of Erlangen-Nuremberg, Cauerstr. 11, 91058 Erlangen, Germany,
christoph.thurner@fau.de, alexander.martin@fau.de

Abstract. In Germany, and in particular in Bavaria, the energy transition towards a more sustainable energy system is an important issue. The nuclear phase-out until 2023 is enacted and the extension of renewable energy sources takes place faster than expected. Optimization models can help to find the optimal extension path for renewable and conventional energy sources from an investment cost perspective. In addition, simulation models can be used to analyze how the electricity demand could be covered by different energy sources in a higher time resolution. In this paper we describe our optimization and simulation framework for the energy transition in Bavaria to perform an energy system analysis. Additionally, we present a coupled approach to analyze the annual energy balances for an optimal extension path.

Key words: Electricity transition, Energy System Analysis, Hybrid Simulation, Energy optimization, Unit commitment problem

1 Introduction

The resolution of the German government phasing out nuclear energy provides an immense challenge for the Bavarian electricity generation system. The share of electricity generated by nuclear power plants in total electricity consumption of Bavaria is about 50 %. According to the Bavarian energy concept “Energie innovativ” [1] the loss of nuclear energy should be compensated by expanding renewable energy sources (RES) significantly faster, building new gas power plants, expanding the grid, and building electricity storages. For instance, the share of RES in total electricity consumption of Bavaria rose to 28.5 % in 2011. In eight years, 50 % of the Bavarian electricity demand should be covered by RES. In order to investigate the effects of the transition of the energy system as a whole, a research project[3] has been launched to develop a combined optimization, simulation, and electricity net model to perform an energy system analysis for Bavaria. The aim of the optimization model is to compute an optimal capacity extension plan for RES and fossil power plants in order to meet the governments targets as well as to ensure maximum security of energy supplies for the planning period until 2023 when all nuclear power plants will be turned off. The goal of the simulation model is to present the most relevant parts of the energy system, to investigate various scenarios, and to study the effects of these scenarios on the future energy balances, the environment, and the electricity imports and exports in a very high time resolution.

In [2] the authors studied 37 tools that can be used to analyze the integration of renewable energy into various energy-systems under different objectives (e.g. availability, type, or geographical area). Connolly *et al.* came to the conclusion that there is no tool available that addresses all issues with respect to the integration of electricity generated by RES. Due to the phase-out of nuclear energy and the building of new power plants, the situation in Bavaria is even more complex.

In this paper we describe our optimization and simulation framework and contribute a coupled approach for one basic scenario to analyze the annual energy balance. The benefit of this approach is to investigate a cost-minimal extension path for RES for a planning horizon of ten years in a high time resolution.

[3] The research project is funded by a consortium of the Bavarian government and various companies from the energy sector and is coordinated by Bayern Innovativ (`http://www.bayern-innovativ.de/cluster-energietechnik/systemanalyse_bayern`) and its Cluster Energy Technology. The optimization model is developed by the Chair EDOM, the simulation model is developed by the Chair of Computer Networks and Communication Systems, and the electricity net model is developed by the Chair of Electrical Energy Systems.

2 Optimization Framework

The optimization framework is a two-stage mixed integer programming (MIP) approach. In order to find a cost efficient capacity extension plan, we implemented an iterative algorithm using the Python-API of Gurobi 5.5 [3], which is a standard solver for MIPs. This idea is an extension of Benders decomposition methodology [4], which has already been analyzed by Nicolosi [5].

The contribution of the framework described in the following is that we do not analyze typical days in the unit commitment sub problems but regard the whole planning horizon with locally refined time spaces. Similar work based on aggregation algorithms has been studied by, e.g. Newman and Kuchta [6] in the field of long-term production planning at underground mines or Hallefjord and Storoy [7] for general integer programming problems.

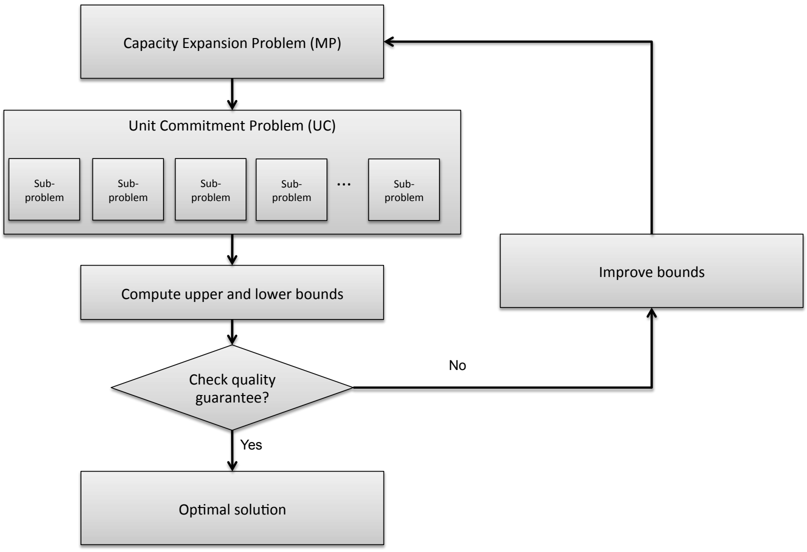


Fig. 1. Representation of the iterative procedure of the implemented decomposition framework and the consecutive refinement of the UC planning horizon.

Fig. 1 shows the iterative procedure of the framework. First we solve the capacity expansion master problem (MP) in the top box in Fig. 1, which is optimized based on the results of the second stage unit commitment problem ensuring to meet the demands on the energy system like, e.g. peak load coverage or system flexibilities. Hence we derive an optimization problem, whose solution guarantees cost minimal investment decisions. Therefore we define the objective function as the sum of discounted annuities of specific investment costs and further fix operation and maintenance (O&M) costs. In the second stage we solve the unit commitment problem (UC), which assures a cost minimal control of the energy system with respect to the installed capacities derived from the MP. The main target of this second stage problem is to ensure security of energy supply with respect to the technical constraints of the energy system. This issue becomes more and more challenging with rising share of RES. This subject has also been analyzed, e.g. by Grimm [8]. Hence the main task is to provide a minimal cost control of the conventional power plants covering the residual load. Note that the commitment order of the conventional plants is according to the merit order, which is based on the marginal costs of the power plants. Hence the objective function is defined as the sum of the variable costs, which consist of variable O&M costs, CO_2 emission costs and fuel costs.

In order to handle the long planning period efficiently, we aggregate the time steps and refine them iteratively at critical points in time. Starting with an initial solution of the MP we solve the UC with a first aggregation of the planning horizon. Consecutively a sequence of MPs over increasingly fine-grained representations of the original UC is solved. In each step, sub problems of the UC are solved. A sub problem is either feasible with respect to an operation schedule, or returns a directive where to refine the representation of the planning horizon in the subsequent iteration. This procedure is sustained until all sub problems are feasible and the convergence check of the optimal solution satisfies a predefined bound.

3 Simulation Framework

Our simulation framework is built to answer different questions about the future electricity generation system of Bavaria. We implemented a simulation framework that includes many system relevant parts on a high level. Therefore we use a hybrid modeling approach which is explained in detail in [9]. We use discrete event modeling for features which change at discrete points in time, e.g. annually changing fuel prices or breakdowns of conventional power plants. Otherwise, the electricity demand is a continuous process with respect to time. Therefore, we use the system dynamic modeling methodology. For realization purposes we use the commercial simulation tool AnyLogic[4].

[4] XJ Technologies Company Ltd. 2012 (`www.xjtek.com`)

In general, the simulation time begins on 1/1/2010 and ends on 1/1/2024. For the years 2010 and 2011 the comparison with official data can be used for validation purposes. Fig. 2 depicts the input and output parameters of our simulation framework. The input parameters on the left side are identical for all scenarios. For instance, the electricity demand is based on load profiles for Germany in an hourly resolution published by the ENTSO-E [10], and feed-in data of photovoltaic systems and wind energy plants provided by the transmission system operators ([11] - [14]) is used to model the feed-in structure of them. Moreover, we take also political conditions such as the German renewable energy law or the phase-out of nuclear power plants into account. Annually changing costs of fuel and CO_2-certificates are used to plan the operation of conventional power plants.

The right side of Fig. 2 shows basic conditions for different scenarios and output parameters. For instance, the extension targets for renewable energy sources can be adjusted to 40 %, 50 %, or 80 % of the total annual electricity consumption. Furthermore, we can investigate the effects of building new gas power plants or to expand the net transmission capacities. The electricity consumption of Bavaria is assumed to stay at 85 TWh by default which is based on the idea that increasing power demand can actually be compensated by increasing efficiency of, e.g. electric devices.

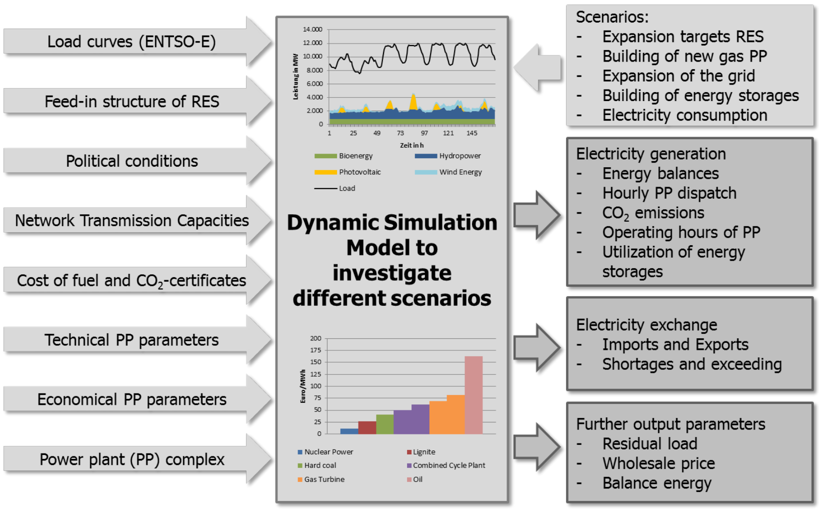


Fig. 2. Overview of input and output parameters.

The investigation of the electricity generation includes, among others, the energy balances on different time scales (annually, monthly, and daily) and the hourly operation plans for conventional power plants, which are needed for the coupling with the electricity net model. Further output parameters are the development of the residual load – defined as the difference between the electricity load and the amount of electricity generated by RES – or the wholesale price.

4 Coupling of the Models and Results

The optimization model is used to determine an optimal capacity extension plan for RES and gas power plants with respect to technical and economic constraints. The cost-minimal extension paths can be used as input parameters for the simulation model and can than be analyzed in more detail. Table 1 shows the optimal extension paths for different RES. We find that additional RES capacities are extended essentially in the second part of the planning period. This is due to fact that the first nuclear power plant is shut off in the end of 2015 and hence capacity shortcomings do not arise until then. Basically cost optimal extensions are added whenever there is a capacity gap. In addition to RES it becomes necessary to build new gas power plants in order to satisfy the increasingly fluctuating residual demand caused by the growing share of RES. Hence the optimization model suggests to build a big 800 MW gas power plant in 2019 and another one in 2022, which will provide enough flexibility to cover the peak residual load and capacity bottlenecks. Fig. 3 depicts the simulated annual electricity generation balance for the years 2014 to 2023 based on the optimized extension paths for RES. The electricity consumption of 85 TWh can be covered in every year. As early as 2021, electricity generated by RES has a share of 50 % of the annual electricity balance. In 2023, the final phase-out of nuclear energy is mainly compensated by electric-

Table 1. Optimal annual extension paths of RES (installed power in MW)

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Photovoltaics	9287	9287	9287	9287	14707	14707	14707	14707	14707	14707
Wind Energy	868	868	868	868	868	868	868	3586	3586	3586
Geothermal	4	4	4	275	275	275	275	275	275	275
Bio Energy	1099	1099	1099	1099	1099	1099	1099	1518	1518	1518
Hydropower	2993	2993	2993	2993	2993	2993	2993	3148	3148	3148

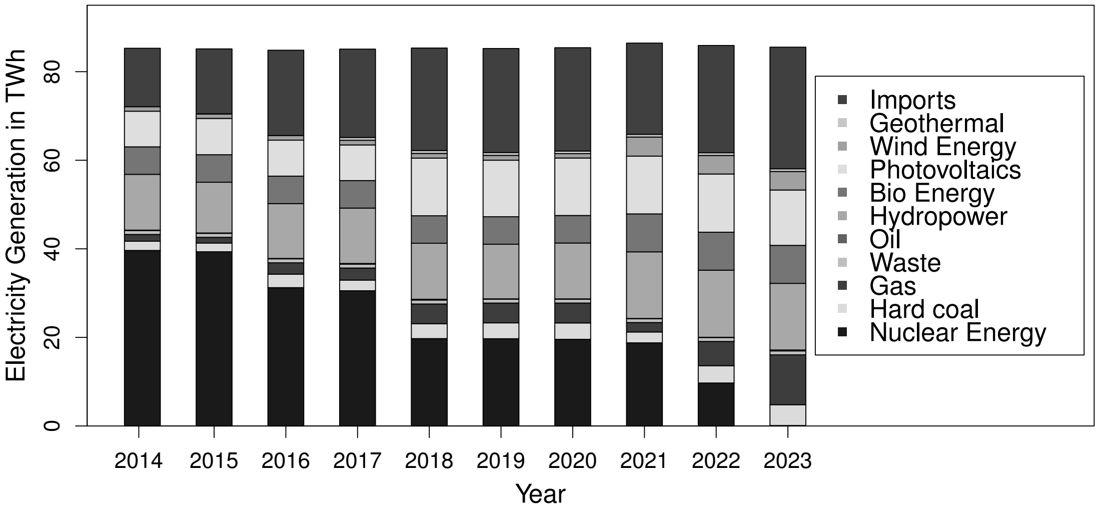


Fig. 3. Annual electricity generation balance.

ity generated by RES (50 %), electricity imports (32 %), and gas power plants (7 %). Further possible outcomes such as higher time resolutions, CO_2-emission balances or full-load hours of new built gas power plants cannot shown due to space limitations.

5 Conclusion

In this paper we present an optimization and simulation framework for the energy transition in Bavaria. With this contribution it is possible to investigate a cost-optimal extension path with the simulation model in detail. Optimization results show that capacities are extended in the second part of the planning horizon, whereas simulation results prove the security of electricity supply. Future work includes the iterative coupling of both frameworks and the coupling with the electricity net model.

References

1. Bayerische Staatsregierung: Bayerisches Energiekonzept “Energie innovativ”, `http://www.energie-innovativ.de/energie-versorgung/` (2011)
2. Connolly, D., Lund, H., Mathiesen, B.V., Leahy, M.: A review of computer tools for analysing the integration of renewable energy into various energy systems. Applied Energy 87 (4), 1059–1082 (2010)
3. Gurobi Optimization. Gurobi optimizer reference manual, `http://www.gurobi.com` (2013)

4. Benders, J.F.: Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik 4(1), 238–-252 (1962)
5. Nicolosi, M.: The Economics of Renewable Electricity Market Integration – An Empirical and Model-Based Analysis of Regulatory Frameworks and their impacts on the power market. Dissertation. University of Cologne (2011)
6. Newman, A.M., Kuchta, M.: Using aggregation to optimize long-term production planning at an underground mine. European Journal of Operational Research 176(2), 1205-–1218 (2007)
7. Hallefjord, A., Storoy, S.: Aggregation and disaggregation in integer programming prob-lems. Oper. Res. 38(4), 619–623 (1990)
8. Grimm, V.: Einbindung von Speichern für erneuerbare Energien in die Kraftwerkseinsatzplanung – Einfluss auf die Strompreise der Spitzenlast. Dissertation. Ruhr-Universität Bochum (2007)
9. Pruckner, M., German, R.: A Hybrid Simulation Model for Large-Scaled Electricity Generation Systems. In: Proceedings of the 2013 Winter Simulation Conference, pp. 1881–1892, Washington (2013)
10. ENTSO-E: Hourly load values of specific country of specific month, `http://www.entsoe.eu/resources/data-portal/consumption/` (2013)
11. TenneT TSO: Network figures, `http://www.tennettso.de/site/en/Transparency/publications/network-figures/overview` (2013)
12. 50Hertz: Grid Data, `http://www.50hertz.com/en/Netzkennzahlen.htm` (2013)
13. Amprion: Grid data, `http://amprion.de/en/grid-data` (2013)
14. TransnetBW: Key Figures RES, `http://www.transnetbw.com/key-figures/renewable-energies-res/` (2013)

Hybrid Simulation Framework for Renewable Energy Generation and Storage Grids

Peter Bazan and Reinhard German

Friedrich-Alexander-Universität,
Erlangen, Germany
http://www7.cs.fau.de

Abstract. Renewable energy sources replace the conventional energy sources such as nuclear power, coal, oil and gas to an increasing degree. As a consequence, fluctuating and decentralized energy production will make up a large part of the energy system. Households and commercial operations can be equipped with photovoltaic panels to generate electricity. We present an improved version of a hybrid simulation framework for renewable energy generation and storage grids. As an example, a model of a gas station micro grid combined with photovoltaics is analyzed.

Key words: Renewable Energy Generation, Storage Grid, Hybrid Simulation, Micro Grid, Energy System

1 Introduction

In the year 2011 renewable energy already covered twenty percent of the final electricity consumption in Germany. It is predicted to reach at least thirty-five percent coverage until the year 2020. Photovoltaics (PV) delivered sixteen percent of the electricity needed. PV are either installed on rooftops or in solar parks and increase the contingent of decentralized energy production. Considering decreasing PV feed in tariff, for energy consumers who are also energy producers (so called prosumers), it is worthwhile to achieve energy self-sufficiency. Due to the fluctuating energy production with PV, energy storages are required.

For the design and analysis of such systems, methods and tools have to be provided. We present a component-based hybrid simulation framework for renewable energy generation and storage grids. It is an advanced version of the previous framework [1], which used the concept of system dynamics (SD) for energy flows and discrete event models for control decisions and fluctuations such as weather and load. The components, like PV, battery, demand, or weather, were coupled by energy flows allowing a uniform interface for each component. In the new version control messages can be interchanged among components enabling more intelligent micro grid control algorithms. This facilitated a redesign yielding a simplified energy flow interface.

Furthermore we present a model of a gas station with PV and a battery – both constructed and analyzed with the presented hybrid simulation framework. The remainder of the article is organized as follows: In Chapter 2 we discuss the related work. Chapter 3 presents the new interface concept of the simulation framework. A micro grid example which is modeled and analyzed using this framework is explained in Chapter 4. Chapter 5 concludes the paper.

2 Related Work

The solar model component is based on the solar model in [2]. The authors present a hybrid simulation model for PV and batteries, which is implemented in AnyLogic [3]. A component-based design of homes is described in [4] that served as a guide for the component-based structure of the presented simulation framework. In contrast to our framework, the authors have no integrated renewable energy. In a more recent version [5] the export of energy is possible, but also there is no PV integrated in a household.

Instead of PV as a renewable energy source, a wind turbine is used in [6]. The model is implemented using AnyLogic, where it is part of a micro-grid simulation with households. There, however, no details about the houses or their components are given. In [7] a refined wind model is used. The integration of wind power would be an interesting extension of the presented simulation framework. A house model created with the previous version can be found at the website [8].

3 Hybrid Simulation Framework

In this chapter we describe the revised version of the component interface for the improved hybrid simulation framework. The framework allows modeling and analyzing renewable energy generation and storage grids. The description of the previous interface design can be found in [1].

3.1 The Communication Interface

The new communication interface *Com* (Fig. 1) of the revised version of the simulation framework is used for the discrete exchange of parameter, internal state, and control information. The *Local Net* in Fig. 1, e.g., receives parameters and control information from the component *Battery*, whereas it controls the *Battery* by sending control messages. Important messages are the amount of charge/discharge power, the local net demands from the battery, the charging state, or the capacity. Now the controller can use more state information of the connected components compared to the previous version of the simulation framework and therefore more sophisticated control algorithms can be applied. In addition, the message oriented interface can be used to insert communication components for the analysis of the performance or the functional behavior of communication technologies and their impact on smart grids.

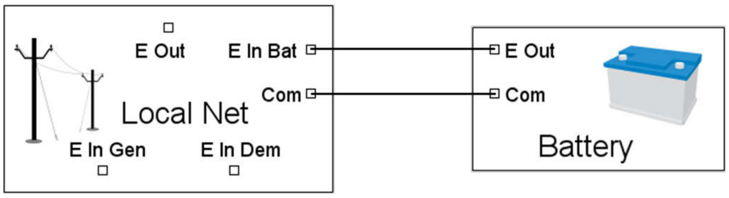


Fig. 1. Two components with communication interface

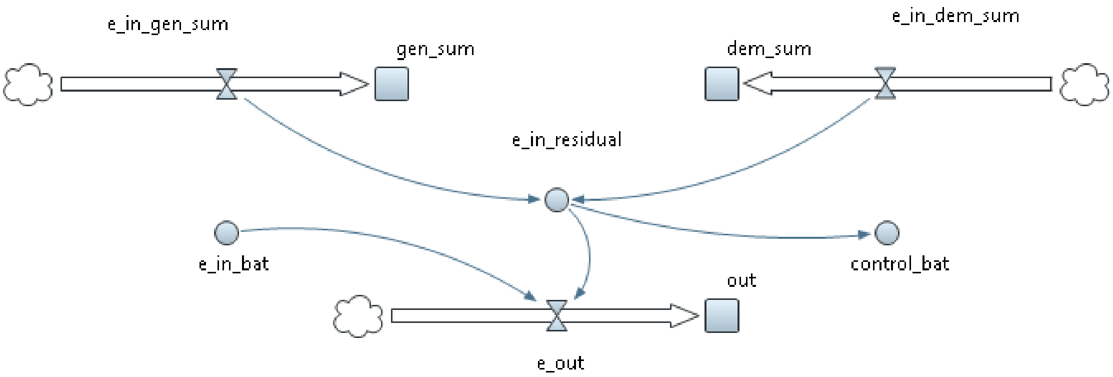


Fig. 2. SD-model of the local net

3.2 The Energy Flow Interface

The implementation of the communication interface facilitates a simplified design of the continuous energy flow interface. For example in the old design, if there were several batteries connected to the local net, the flow towards them had to be split with respect to the charge state and the maximal power in- and output of each individual battery. Now, this information is processed via the new communication interface. Therefore the continuous energy flow interface can be simplified by allowing only positive or negative flows from components on a hierarchically lower logical level to components on a hierarchically higher logical level. These flows are implemented by SD-models and the SD-models of the components are connected at the start-up of a simulation run by exchanging connection messages.

In Fig. 1 the hierarchically lower level component is the battery. Hence, the energy flows from the connection *E Out* of the battery to the connection *E In Bat* of the local net. *E In Bat* is connected to *e_in_bat* of the SD-model of the local net (Fig. 2), *E In Gen* is connected to *e_in_gen_sum*, *E In Dem* is connected to *e_in_dem_sum*, and *E Out* to *e_out*. The local net calculates the energy balance from simply adding all positive energy flows from *E In Gen* and all negative energy flows from *E In Dem*. This value is used for the computation of the charge or discharge power of the battery. The resulting charge/discharge power is sent over the communication connection. The controller of the battery then adjusts the energy flow of the battery connection *E Out*. The sum of the flows *E In Gen*, *E In Dem*, and *E In Bat* is finally the positive or negative flow of *E Out* for the power network.

4 Renewable Energy Generation and Storage Grid

With the improved simulation framework, a model of a gas station is constructed (see Fig. 3). The station is a tank farm without office or convenience store. The energy users are the pumps, the payment system, and the illumination. The model consists of two fields of photovoltaic panels which are connected to a stochastic weather module. In addition to that, each of the panel is equipped with a DC/DC converter. The gas station has five pumps (Fig. 4) which are controlled by the gas up controller using the new communication interface. In the more energy flow oriented old version of the framework, no interface was provided for such control messages.

The gas up controller is a model of the traffic at the gas station. The pumps are connected to the local net by DC/AC converters. Both, the DC/DC and DC/AC converters are models reflecting their respective efficiencies. The local net cal-

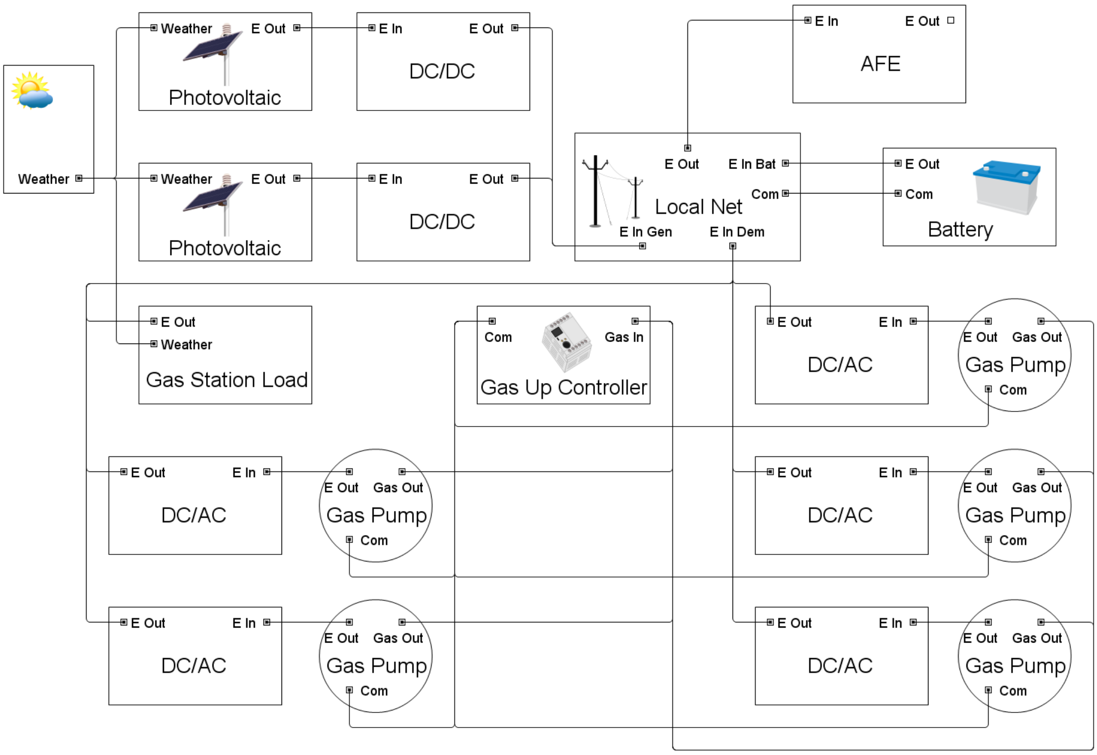


Fig. 3. Model of the gas station with a field of two PV systems and five gas pumps

culates the energy balance and stores excess energy in the battery. If this is not possible (the battery is fully charged or the maximum charging power is reached), the excess energy is exported via the active front end (AFE) to the net. If the local net calculates an energy demand, the energy is taken from the battery. Whenever the battery is fully discharged, the energy has to be imported from the net. The interactive graphical user interface of the model is shown in Fig. 5.

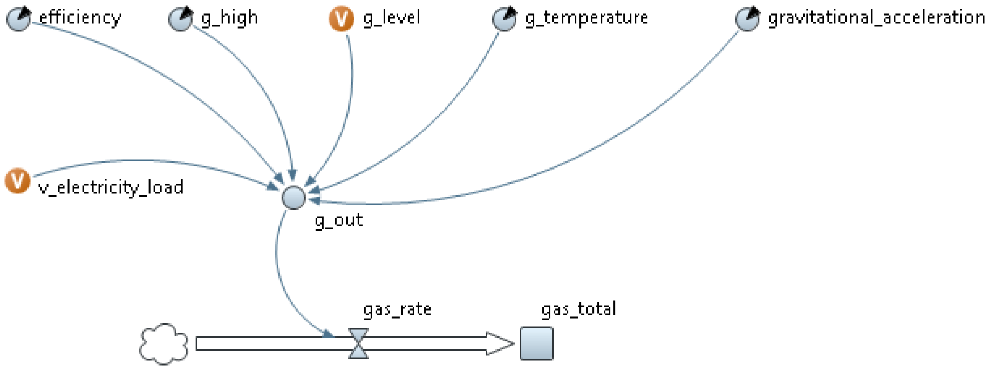


Fig. 4. SD-model of the gas pump component

The simulation period for the gas station model has been set to one year and one hundred simulation runs were performed for each experiment. In Table 1 the results of the simulation runs of the gas station model for different battery capacities are given. It shows the mean values per year for the electricity generation of the PV, the electricity demand of the pumps, payment system, and illumination, the electricity exported to and imported from the outer power supply net, and the electricity loss due to the converters and the battery.

With a PV peak power of 6 kW, the mean yearly electricity production is 4.9 MWh. With no battery (battery capacity set to 0 kWh) the mean energy consumption of the gas station's energy users was about 4.1 MWh. Together with the energy loss of the converters, the gas station needs 4.3 MWh. Therefore the generated electricity serves the energy need of the gas station, but nevertheless, 3.9 MWh have to be exported and 3.3 MWh have to be imported from the power supply net. The generation and the consumption of electricity do not match in time.

Table 1. Mean electricity generation (PV), electricity demand (pumps, payment system, and illumination), electricity exported to the outer power supply net, imported energy, and electricity loss (converters and battery) per year of the gas station with different battery capacities

Batt. Capacity [kWh]	Mean Generation [MWh]	Mean Demand [MWh]	Mean Export [MWh]	Mean Import [MWh]	Mean Loss [MWh]
0	4.9	4.1	3.9	3.3	0.27
10	4.8	4.1	1.1	1.1	0.73
20	4.9	4.1	0.8	0.8	0.86
30	4.9	4.1	0.6	0.6	0.95

A battery can store excess electricity, increasing the energy self-sufficiency of the gas station. This effect is shown by three experiments with different capacities of the battery (Table 1). With a battery capacity of 10 kWh and a PV peak power of 6 kW, the gas station's mean export and mean import of electricity is reduced to 1.1 MWh. The PV generates just the amount of electricity needed for the demand and the energy loss of the converters and the battery. Increasing the battery size to 20 kWh or 30 kWh results only in a slight decrease of the exported and imported electricity.

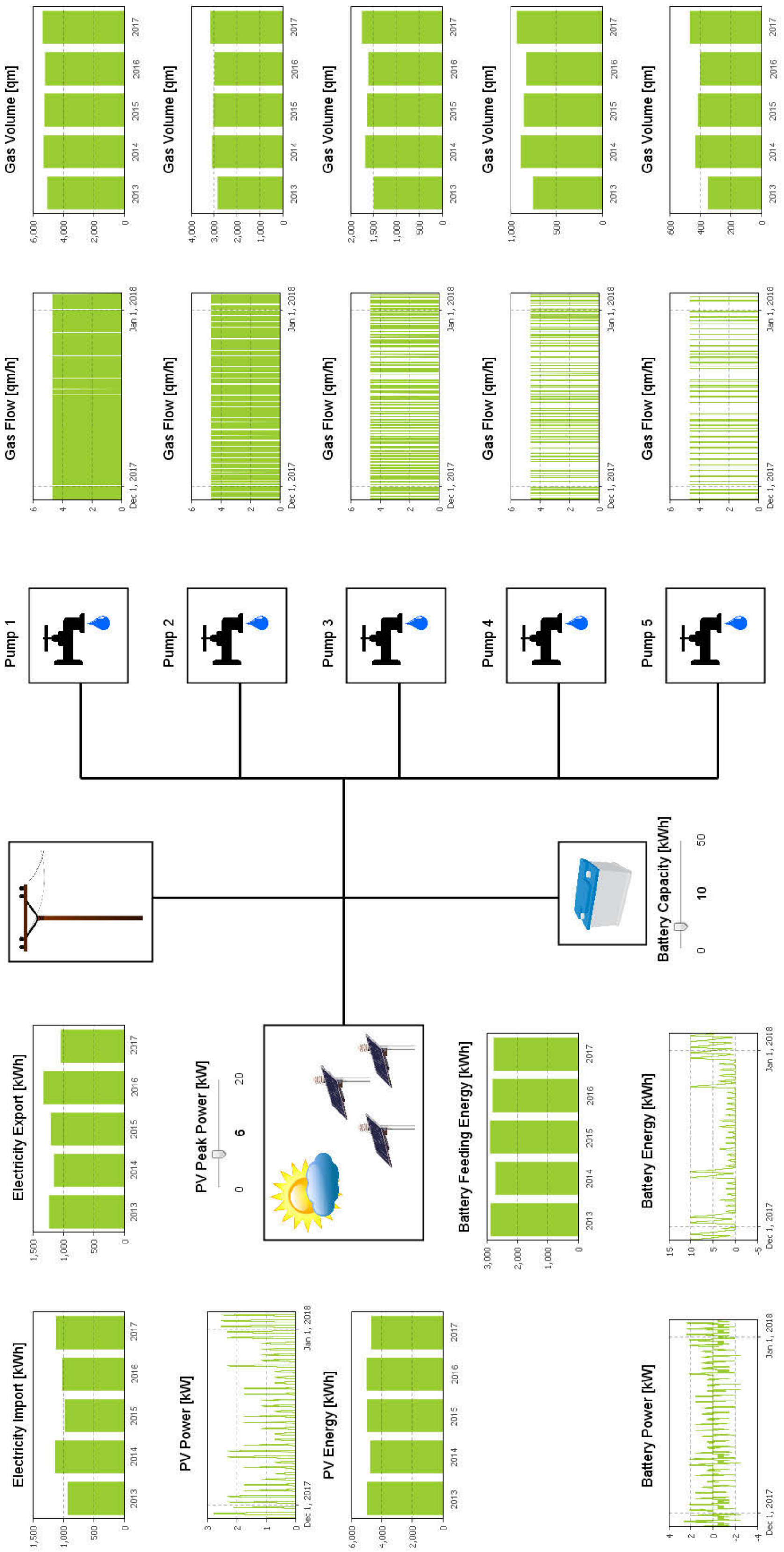


Fig. 5. Interactive graphical user interface of the gas station

5 Conclusion

In this article we presented an improved version of a hybrid simulation framework for the modeling and analysis of renewable energy generation and storage grids. For this purpose the interface was redesigned, such that control information is exchanged via communication ports. Given that, the entire system may be designed by implementing the energy flows from components on a hierarchically lower level to components on a hierarchically higher level. With the improved simulation framework a gas station with PV and battery was modeled and analyzed with the help of the implemented interactive graphical user interface. For the chosen parameter sets, the possibility of a balanced energy import and export of the gas station was revealed.

References

1. Bazan, P., German, R.: Hybrid Simulation of Renewable Energy Generation and Storage Grids. In: Proceedings of the 2012 Winter Simulation Conference pp. 237:1–237:12. Berlin (2012)
2. Mazhari, E. M., Zhao, J., Celik, N., Lee, S., Son, Y.-J., Head, L.: Hybrid Simulation and Optimization-Based Capacity Planner for Integrated Photovoltaic Generation with Storage Units. In: Proceedings of the 2009 Winter Simulation Conference pp. 1511–1522. Austin (2009)
3. XJ Technologies Company Ltd.: AnyLogic, `http://www.anylogic.com`
4. Molderink, A., Bosman, M. G. C., Bakker, V., Hurink, J. L., Smit, G. J. M.: Simulating the Effect on the Energy Efficiency of Smart Grid Technologies. In: Proceedings of the 2009 Winter Simulation Conference pp. 1530–1541. Austin (2009)
5. Bakker, V., Molderink, A., Bosman, M. G. C., Hurink, J. L., Smit, G. J. M.: On Simulating the Effect on the Energy Efficiency of Smart Grid Technologies. In: Proceedings of the 2010 Winter Simulation Conference pp. 393–404. Baltimore (2010)
6. De Durana, J. M. G., Barambones, O.: Object Oriented Simulation of Hybrid Renewable Energy Systems Focused on Supervisor Control. In: Proceedings of 12th IEEE International Conference on Emerging Technologies and Factory Automation pp. 1–8. Palma de Mallorca (2009)
7. Kremers, E., Lewald, N., Viejo, P., De Durana, J. M. G., Barambones, O.: Agent-Based Simulation of Wind Farm Generation at Multiple Time Scales. In: Wind Farm - Impact in Power System and Alternatives to Improve the Integration Intech (2011)
8. Simulation framework i7-AnyEnergy, `www7.cs.fau.de/energy`

Towards Simple Models for Energy-Performance Trade-Offs in Data Centres

Boudewijn R. Haverkort and Björn Postema*

University of Twente
CTIT Centre for Dependable Systems and Networks
b.r.h.m.haverkort@utwente.nl, b.f.postema@utwente.nl
http://www.utwente.nl/ewi/dacs/

Abstract. In this paper we advocate the use of simple stochastic models to analyse the energy-performance trade-off in data centres. Recently such trade-offs have received increased attention, however, the tools used to make such trade-offs are largely based on simulation and real-life experiments. Although simulations studies are very helpful, we think that simple analytical models, or models based on stochastic Petri nets (or similar description techniques) can be very fruitful in guiding design processes in the early phases. Similarly, we do think that experimental work is very important, however, its results come "after the fact" in the sense that the system has been built already once the experiments are being performed. Our claim is that the use of simple models early in the design phase provides a very good return on investment. This short paper presents some preliminary models that can be used for early-in-design trade-off analyses.

Key words: Data centres, energy, performance, analytical models, matrix-geometric methods, stochastic Petri nets.

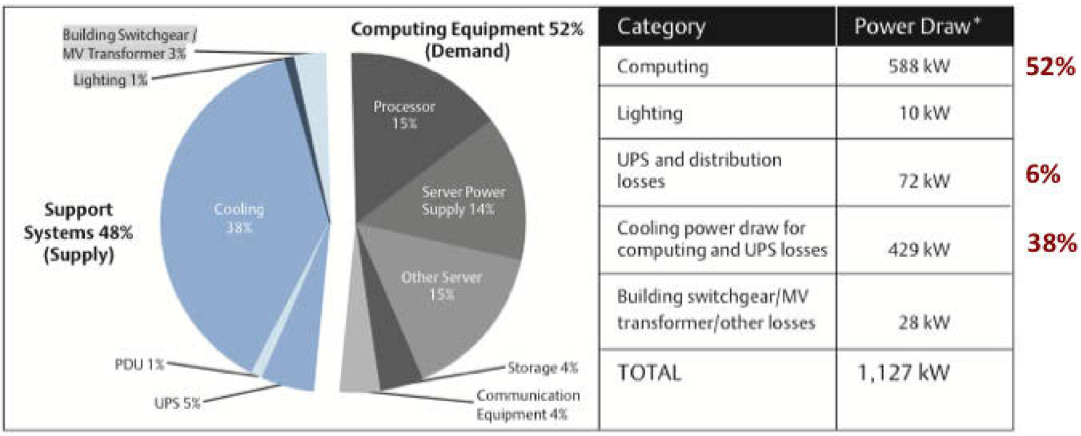
1 Introduction

Recent studies have revealed that ICT equipment consumes 2-3% of all electrical energy, and the trend is that this number is growing [13]. Although other sectors

* The work in this paper has been supported by the Dutch national STW project *Co-operative Networked Systems*, as part of the program "Robust Design of Cyber-Physical Systems".

are much more energy hungry, i.e., traditional industries amount for about 40%, transportation for about 20%, and residential use for about 12%, still it is important to make ICT less energy hungry. A recent EU forecast states that in 2020 93 TWh is used for ICT, which is the equivalent of 106 million 100 Watt light bulbs burning continuously for a year. Data centres are particularly energy hungry: per 60×60 cm^2 floor usage in a data centre, the annual CO_2 emission is as much as 1200 kg. Of course, this is no news any more; many project have in the meantime been started to investigate “green ICT”.

Within a data centre, around 50% of the energy is being used for ICT, the remaining part being used for lighting, UPS (uninterruptible power supply), cooling, etc., see Figure 1. Furthermore, it is well-known that energy use for ICT in a data centre is multiplied through the so-called *cascade effect* [7], which states that a 1 Watt saving in CPU power, leads to 0.18 W reduction for DC-DC conversion, 0.31 W reduction in AC-DC power conversion, 0.04 W in power distribution, 1.4 W in UPS, 1.07 W in cooling, and finally 0.1 W in power reduction for switchgear and transformer capacity. In total, this yields a factor 2.84. This is both good and bad news: more energy use for ICT implies more energy use elsewhere (bad news), however, the good news is that savings at the ICT are also translated into savings elsewhere.



Category	Power Draw*	
Computing	588 kW	52%
Lighting	10 kW	
UPS and distribution losses	72 kW	6%
Cooling power draw for computing and UPS losses	429 kW	38%
Building switchgear/MV transformer/other losses	28 kW	
TOTAL	1,127 kW	

Fig. 1. Energy use within a data centre (picture from [4])

In this paper we advocate the use of simple stochastic models to analyse the energy-performance trade-off in data centres. Recently such trade-offs have received increased attention, however, the techniques used to make such trade-offs are largely based on simulation and experiments. Although simulations studies are very helpful, we think that simple analytical models, or models based on stochastic Petri nets (or similar description techniques) can be very fruitful in guiding design processes in the early phases. Similarly, we do think that experimental work is very important, however, its results come "after the fact" in the sense that the system has been built already once the experiments are being performed. Our claim is that the use of simple models early in the design phase provides a very good return on investment.

Below, we will first sketch the bigger picture in energy savings for data centres, followed by a simple model-based approach that will allow us to investigate design trade-offs regarding energy and performance, thereby using simple Markovian models based on stochastic Petri nets.

2 Possible energy reduction steps

In this section we briefly discuss a number of ways to decrease energy use in data centres [7], that range from very practical ("moving boxes") to more advanced (requiring elaborate sensory equipment) and more software-oriented.

1. **Data centre ICT equipment**:
   - The use of *low voltage processors* will directly decrease the power usage with some 30%, thereby not necessarily impacting performance, although detailed studies have to be made to ascertain that.
   - The use of *high-end power supplies*, with efficiency 90% instead of the typical 70% that is standard for low-end power supplies in consumer computing equipment. Furthermore, the power supplies should be chosen such that they, under normal load circumstances, operate at their optimum.
   - The use of *blade servers*, that use multiple processing boards with shared IO, fans, and power supply.
   - The use of emerging techniques for *green networking*; in the current data centre literature there appears to be a focus on computing only.
2. **Data centre (power) management software**:
   - The use of *advanced power management software*, that make that servers or server groups can be switched off completely while still meeting the performance requirements; this will be elaborated upon in the next section.

   - Advanced *server virtualisation* software can be used to increase server utilisation and to reduce the number of active servers.

3. **Data centre power supply**:
   - Higher voltage AC *power distribution* within a data centre can decrease overall power usage, as higher voltage transport is more efficient than low voltage transport; the EU is doing better here (with their standard 240V than the US with standard 110V).
   - The use of *more efficient UPS systems*, that do avoid the double conversion, from the external AC source, to the DC storage and buffering, and the AC end-use.
4. **Data centre cooling**:
   - The use of better *spatial arrangement of servers* and cooling (use of hot/cold aisles) and higher room temperatures (28 vs. 20 degrees Celsius) and *variable capacity precision cooling* (instead of simple overall room cooling) to cool just there where it is needed.
   - *Per server/system monitoring and control* of temperature, humidity, etc., to further increase cooling efficiency; this requires the installation of an advanced (wireless) sensor system.

Of the above energy reduction steps, we will focus below on an approach that allows for the analysis and effect of dynamic power management software, server virtualisation and per-server monitoring, as these lend themselves very well for a model-based approach.

3 A model-based approach

As stated in the introduction, most work on data centre energy efficiency focuses on simulation and experimentation. The focus in this short paper is on analytical and numerical techniques. Recent work along these lines is still rather limited, but interesting (and partly similar) work on using Markov chains can be found in [8,10,14]. Another interesting approach based on stochastic Petri nets is [3], in which the effect on the energy usage of on-demand creation and deletion of virtual machines is studied.

The basic idea of our models is illustrated in Figure 2: a data centre serves a job stream from the outside world, buffers the incoming jobs and subsequently schedules and executes them, on the basis of the job requirements and the system-internal state information that is available. What that state information exactly is, largely depends on the data centre: it might involve only information on job queue lengths or server utilization, but can also include information on temperature and humidity in parts of the data centre, or information in networking bottlenecks, to give just a few examples.

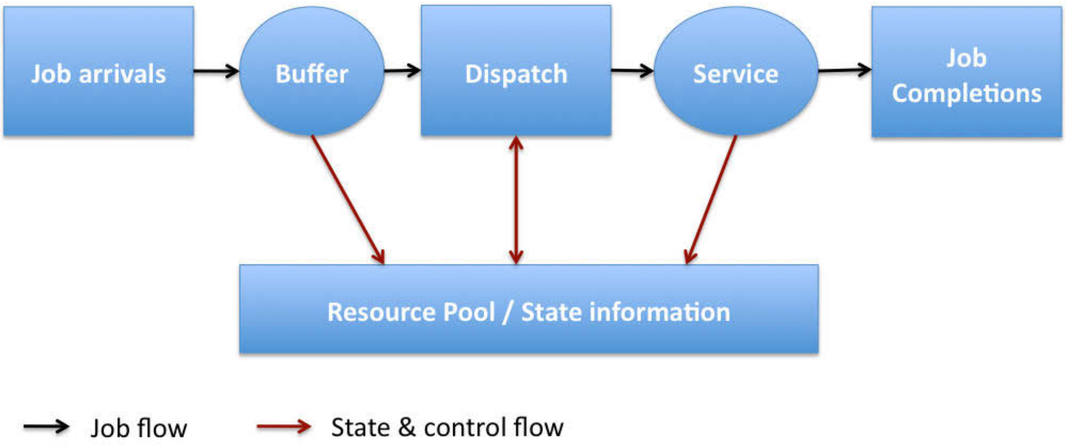


Fig. 2. The basic model for energy/performance trade-offs in data centres

Of course, the model in Figure 2 cannot be used for any computation; it has to be made more concrete, e.g., in the form of the simple (infinite-state) stochastic Petri net [16,17] given in Figure 3. In this model, jobs arrive according to a Poisson process (transition `Arrivals` with rate λ) and are buffered in place `Buffer`. If the server is switched off upon arrival (token in place `Off`), it has to be switched on first (via transition `Setup`), leading to an extra delay (transition `Delay` with rate α) and extra energy use (as long as there is a token in place `Booting`) before actual processing (token in place `Processing` can starts (transition `Service`). Once the processing finishes, with rate μ, the server is moved to place `Idle` where it will not stay when there are other jobs buffered (since transition `Start` will fire immediately in that case). If there are no other jobs waiting to be served, the server will stay idle for some amount of time, before it is switched to a lower-energy state (via transition `Release`). The time-out value (exponential rate β of transition `Release`) in relation to the setup delay (exponential rate α) as well as the energy usage parameters (non-zero reward rates "`E`" for places `Booting`, `Processing`, and `Idle`), allow for making a trade-off between energy usage and performance requirements. Energy usage is modelled using these *rate rewards,* with the following semantics: as long as there is a token in place, say, `Processing`, then energy is used with rate $E_{\texttt{Proc}} = 200$. Of course, this could be extended to also include *impulse rewards,* that is, impulses of energy being used instantaneously when certain transitions fire.

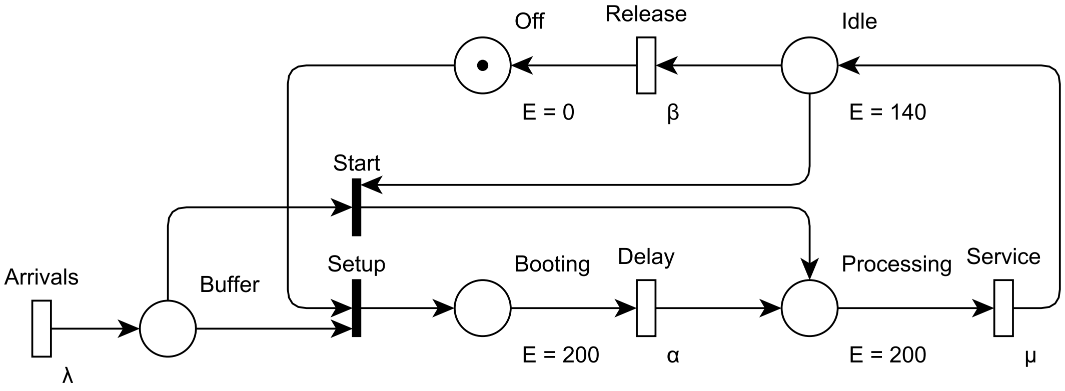


Fig. 3. An infinite-state stochastic Petri net for energy/performance trade-offs in data centres

Note that the above is the simplest possible model, but that it can be easily made more advanced by adding multiple servers, by using more advanced workload models (such as multi-class arrival patterns and multi-class service times), using phase-type distributions, the use of deterministic timing, etc.

Once the model has been specified, it can be solved for either its steady-state behaviour, or its time-dependent (transient) behaviour. In the former case, performance measures such as throughput, mean delay, server utilisation, etc., can be computed, as well as the overall energy-usage rate, that is, the power consumption. In case of a transient analysis, say, for some finite period $[0, t)$, the expected cumulative number of jobs processed and the expected amount of energy used for that can be computed. In its full generality, we might need to use discrete-event simulation to evaluate the model, however, in restricted cases, a numerical analysis can be performed via the automatically generated underlying Markov chain (see [11]). In particular, for the model at hand, we used the Möbius toolset to evaluate the models fully numerically [5].

We now provide an example of the type of trade-offs that can be made. We have used the following parameters (taken from [2,4,9]) for the model: $\lambda = 0.007$ (this is a low rate, but we are only modelling one server here), $\alpha = 0.01$ (giving a mean boot time of 100 s), μ is taken from the set $\{0.01, 0.025, 0.1, 1.0\}$, and the power levels for processing and booting are 200 W, and for idling 140 W. Finally, we let the time-out rate β increase from 0.01 to 1.0 (in steps of 0.01). Figure 4 presents a typical (mean) power-response time trade-off, with on the x-axis the mean response time, and on the y-axis the mean power used, for the four different service rates μ (from bottom-left to top-right: 1.0 (purple), 0.1 (blue), 0.025 (green) and 0.01 (red)). On every curve each cross signifies a different value for the time-out rate β: on the left-most end of each curve the value $\beta = 0.01$ (long average delay before shut-down) and towards the right larger values for β (quicker shut-down); the right-most point is the case $\beta = 0.01$. Clearly, if the time-out delay is, on average, larger (towards the left end of the curves), the server is, upon job arrivals, most often still switched-on but idle, hence, job processing can quickly start (no booting delay), resulting in a lower mean delay, at the cost of higher mean power usage. Conversely, if an idle server is more quickly switched off (towards the right on the curves), a larger mean delay is perceived, but a lower mean power-usage is the gain reached.

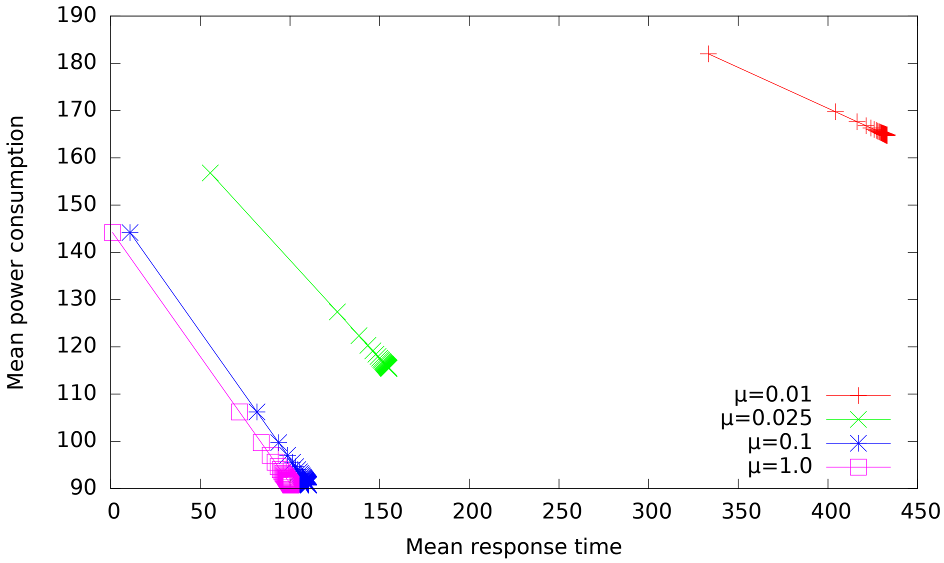


Fig. 4. Mean response time against mean power consumption, using the simple data centre model, for four different service rates

We finish this paper by noting that the model of Figure 3 has as special feature that its underlying Markov chain exhibits a repetitive structure, as depicted in Figure 5, such that efficient matrix-geometric methods can be employed for its solution. In this figure, states of the form (o, n) $(n = 0, 1, 2, \cdots)$ signify states in which the server is idle and there are n jobs in the system, states of the form (p, n) $(n = 1, 2, \cdots)$ signify states in which the server is processing a job and there are in total n jobs in the system, and state $(\mathtt{i}, 0)$ signifies that the server is idle, and that there are no jobs in the system (any more). In three coloured blocks, we indicate the rate rewards to be associated which each of the states. By solving the underlying CTMC, using matrix-geometric methods, the individual steady-state probabilities can be computed, from which mean performance measures such as throughput and delay can be derived. It should be noted that the type of model sketched here, is not new as such. In the mid-1990's, work has been done on connection management in network systems, in which efficient trade-offs had to be found between network connectivity time (hiring of bandwidth) and connection set-up delays; cf. [12,16,17]. The additional parameter of consideration here is energy; the models developed then, can be extended easily towards the needs we have now.

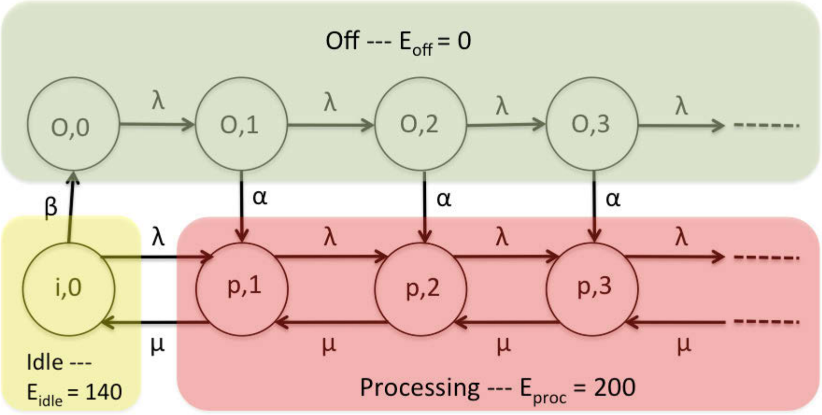


Fig. 5. The underlying infinite-state CTMC for the basic model for energy/performance trade-offs in data centres

4 Conclusion

In this short paper we have proposed the use of analytical and numerical models for the evaluation of, especially, dynamic power management strategies for data centres. We have shown the basic model structure, and given a concrete example of the simplest possible model. Starting from the generic model, many more advanced models can easily be developed and evaluated, using one of the many available stochastic Petri net tools. We also provided a concrete example of the type of trade-offs that can be made.

References

1. T. Bostoen, J. Napper, S. Mullender, Y. Berbers. Minimizing energy dissipation in content distribution networks using dynamic power management. *Proc. Third IEEE Int'l. Conference on Cloud and Green Computing*, pp.203–210. 2013.
2. L.A. Barroso and U. Hölzle. The case for energy-proportional computing. *IEEE Computer* **40**(12): 33–37, 2007.
3. D. Bruneo, A. Lhoas, F. Longo, A. Puliafito. Analytical evaluation of resource allocation police in green IaaS clouds. *Proc. Third IEEE Int'l. Conference on Cloud and Green Computing*, pp.84–91. 2013.
4. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels. Dynamo: Amazon's highly available key-value store. *Proceedings of 21st ACM SIGOPS Symposium on Operating Systems Principles*, pp.205–220, 2007.
5. G. Clark, T. Courtney, D. Daly, D. Deavours, S. Derisavi, J.M. Doyle, W.H. Sanders, and P. Webster. The Möbius Modeling Tool. *Proceedings of the 9th International Workshop on Petri Nets and Performance Models*, IEEE CS Press, 2001, pp. 241-250. See also `https://www.mobius.illinois.edu/`.
6. N. El-Sayed, I. Stefanovici, G. Amvrsiadis, A.A. Hwang. Temperature management in data centres: Why some (might) like it hot. *Proc. ACM Sigmetrics*, pp. 163–174, 2012.
7. Emerson Network Power. Energy Logic: Reducing data centre energy consumption by creating savings that cascade across systems. 2009.
8. A. Gandhi, M. Harchol-Balter. How data size impacts the effectiveness of dynamic power management. *Proc. 49th IEEE Allerton Conference on Communication, Control & Computing*, 2011.
9. A. Gandhi, M. Harchol-Balter, M. Kozuch. Are sleep states effective in data centers? *Third IEEE International Green Computing Conference*, pp.1-10, 2012.
10. A. Gandhi, S. Doroudi, M. Harchol-Balter, A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Proc. ACM Sigmetrics*, pp.153-166, 2013.
11. B.R. Haverkort. *Performance of computer-communication systems: A model-based approach*. John Wiley & Sons, 1998.
12. G. Heijenk, B.R. Haverkort. Design and evaluation of a connection management mechanism for an ATM-based connectionless service. *Distributed Systems Engineering* **3**: 53–67. IEE/IOP Publishing, 1996.
13. J.G. Koomey. Worldwide electricity used in data centres. *Environmental Research Letters* **3**: 1–8, 2008.

14. P.J. Kühn, M. Mashaly. Performance of self-adapting power-saving algorithms for ICT systems. *Proc. IFIP/IEEE International Symposium on Integrated Network Management*, pp. 720–723, 2013.

15. G. Mone. Redesigning the data centre. *Communications of the ACM* **55**(10): 14–16, 2012.

16. A. Ost, B.R. Haverkort. Analysis of windowing mechanisms with infinite-state stochastic Petri nets. *ACM SIGMETRICS Performance Evaluation Review* **26**(2): 38-46, 1998.

17. A. Ost. Performance of communication systems: A model-based approach with matrix-geometric methods. Ph.D. thesis, RWTH Aachen University, Department of Computer Science. Published with Springer Verlag, 2001.

Evaluation of Four Possible Load Shifting Strategies for Electric Vehicles Utilizing Autoregressive Moving Average Methods for Electricity Price Forecasting

Jürgen Wenig and Thorsten Staake

University of Bamberg, Energy Efficient Systems Group
An der Weberei 5, 96047 Bamberg, Germany
http://www.uni-bamberg.de/en/eesys

Abstract. This paper quantifies and compares monetary saving effects that can be achieved by applying different instances of a load shifting system to battery charging of electric vehicles. Along this line, we evaluate four possible load shifting strategies including two methods for predicting energy prices. The proposed strategies refer to a demand side management where electricity customers actively respond to a fluctuation of market prices within 24-hour cycles. We find that forecasting strategies outperform fixed charging times only by a relatively small margin.

Key words: Load Shifting, Load Shifting Strategies, Electricity Price Forecasting, Autoregressive Moving Average Methods

1 Evaluation of Load Shifting Strategies

German policy makers have formulated ambitious targets to accelerate the adoption of electric vehicles: Until the year 2020, more than one million electric vehicles should be registered within the country [1]. While critics argue that these goals are too aggressive and not fully realistic, there nevertheless appears to be some agreement that electric vehicles will be the future of individual road transport and sooner or later overtake convectional cars in numbers. Along this development, systems for battery charging will considerably gain importance. The development goes hand in hand with the fast adoption of renewable electricity sources and the creation of a fluid electricity market in order to handle the partly stochastic electricity supply by rewarding demand flexibility on the side of the consumers.

In this research in progress paper, we aim at optimizing the charging cost of electric vehicles under the assumption that tariffs are available that follow the spot market price for electricity (plus some fixed and/or relative margin for the retailer, the grid operator, etc.). In order to achieve high validity of the results, we use hourly intraday electricity prices from the EEX PHELIX stock market index for Germany and Austria that are available online from the energy exchange. Moreover, we use the specification of an electric vehicle that is commercially available (a BMW i3 with a 125 kW synchronous motor, a usable battery capacity of 18.8 kWh, and a norm-cycle consumption of 12.9 kWh per 100 km) as a model car in order to quantify possible savings. The charging profile introduces an exemplary car owner who uses the car from 07:00 AM to 07:00 PM. The remaining time period can be exploited for a load shifting strategy [2].

The first strategy under observation is rather simplistic and primarily serves as reference scenario. Here, the charging process starts at constant power whenever the test user arrives at home. This strategy is referred to as forward scheduling. The second strategy takes advantage of an inherent characteristic of 24-hour electricity prices, namely the fact that energy prices regularly achieve a daily minimum between after midnight and before early morning, as historic price curves reveal. The car is charged in such a way that the battery is full just in time at 07:00 AM. This strategy is named backward scheduling. The third strategy uses an exponential smoothing technique in order to determine and exploit the possibly cheapest time windows in accordance to the prediction. This strategy substitutes for the forecast of strategy two. The fourth strategy is the most advanced approach as it aims at dynamically determining and exploiting time windows of low energy prices by predicting electricity prices using an autoregressive-moving-average (ARMA) model [3]. We use the Akaike Information Criterion (AIC) as a means for selecting the best model, considering manual data adjustments and seasonality presumptions.

Subsequently, the most promising predictive strategy is selected by comparing the mean absolute deviation (MAD) of predicted values versus actual values. Finally, the financial saving potential is exemplary estimated by comparing both the second and the best of strategies three or four with strategy one.

The study illustrates that basic time series forecasting methods for electricity prices can be used to improve the exploitation of low price time windows on the electricity market. Therefore, owners of electric vehicles, with short loading times in particular, profit from the application of a demand side management of battery charging. The possible savings due to an active reaction to market fluctuations can be approximated, leading to a hands-on recommendation for e-car owners.

The preliminary findings, as illustrated in Fig. 1, reveal that the forecast strategies three and four outperform strategies one and two in terms of low price time window exploitation on a sample day. It shows that the simple approaches in strategies one and two fall short in comparison to strategies three and four.

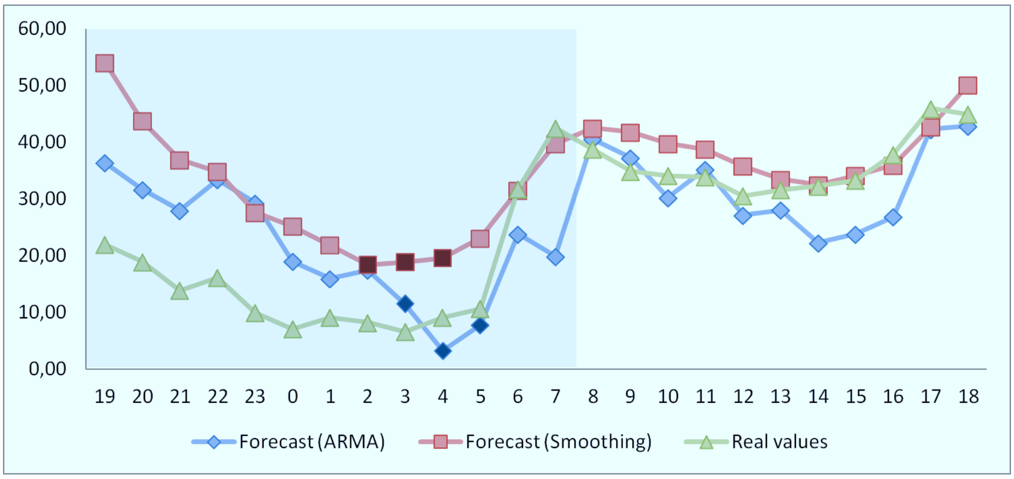


Fig. 1. Forecast 03.11.2013-04.11.2013

For the sample day, overall savings for strategy three amount to 56.3% compared to strategy one and 53.6% compared to strategy two. The application of strategy four saves 51.9% compared to strategy one and 48.9% compared to strategy two.

2 Conclusion and Outlook

Strategies that use forecasting of electricity prices help to reduce the cost of charging electric vehicles. Not surprisingly, these strategies outperform approaches that utilize simple static charging times that are defined according to historic price curves. However, the earnings appear to be rather small – even if one assumes that perfectly accurate price forecasts were available. The reason for the small effect size is the largely self-similar profile of daily electricity prices that allow for defining fixed but reasonably accurate time windows of low prices in combination with the relatively long charging times of the batteries. Nevertheless, one can assume that forecasting strategies will become more attractive if the volatility of electricity prices increases along the growing share of electricity from stochastic energy sources such as wind and solar. On the other hand, forecasting in combination with large loads may reduce the price variability on the spot market, which might reduce the overall effect size. Future work should thus consider more complex models that take the feedback effect into account in order to obtain a more precise picture of the benefits of charging strategies.

References

1. German Federal Government: Nationaler Entwicklungsplan Elektromobilität der Bundesregierung. Berlin (2009)
2. Prüggler, N.: Economic potential of demand response at household level—Are Central-European market conditions sufficient? Energy Policy (60), pp. 487–498 (2013)
3. Contreras, J., Espinola, R., Nogales, F. J., Conejo, A. J.: ARIMA models to predict next-day electricity prices. IEEE Transactions on Power Systems, vol. 18(3), pp. 1014-1020 (2003)

Smart Grid Communication Architecture

Ullrich Feuchtinger, Kolja Eger, Reinhard Frank and Johannes Riedl

Siemens AG
Otto-Hahn-Ring 6
D-81739 Munich, Germany

Abstract. The steeply rising number of sensors and actuators being built into today's power grid, the need for an intelligent processing of the large amounts of data, and the manifold roles of the different Smart Grid stakeholder ask for suitable communication network architectures and technologies to address the Smart Grid application's requirements. This article will summarize how the Smart Grid communication architecture should look like. Based on a first analysis of the Smart Grid applications seen for the next five years some conclusions are drawn on suitable communication network technologies fitting to the different domains of the presented architecture.

Key words: Smart Grid, Communication Architecture

1 Introduction

The Smart Grid is an intelligent, self monitoring and highly automated electric power system that is continuously optimized by selective acting, controlling and adjusting in order to build an improved grid. It is adaptable to new "prosumer" requirements and capable for consistent IT integration [1]. Information and Communication networks that are seamlessly integrated in a Smart Grid Architecture enable the end-to-end data connectivity. Based on this connectivity power generation, transmission, distribution and consumption is facilitated. Beside the traditionally approach that the power generation is based on a small number of large power plants following current consumption it is a key characteristic of the Smart Grid that the current consumption is managed based on the available electrical energy also from distributed and renewable energy resources. The intelligent balancing between Energy Production, Energy Transmission & Distribution, Energy

Storage and Energy Consumption taking grid constraints in transmission and distribution networks into account comprises the Smart Grid scope.

2 Devices of the Smart Grid

Power grids are becoming smarter and getting equipped with many different devices for all kinds of new applications. These devices are sensors (e.g. Smart Meters, Phasor Measurement Units, ...) and actuators (e.g. controllable transformers, protection switches, ...) to collect/understand the status of the power grid and assure power reliability and quality at any point in time. Consequently, the basis for engineering, commissioning and operating Smart Grids is the intelligent connection and use of a huge number of sensors and actuators. Due to the increasing decentralization, the number of involved devices and the complexity of the system increase. Further on, there are many stakeholders like network operators, energy producers and consumers as well as service providers [2] involved in Smart Grids that have to fulfill certain roles regulatory bodies have transferred to them. Thus a suitable networking communication architecture has to be developed for Smart Grids addressing the needs of all stakeholders also.

3 Communication Network Domains

Smart Grid communication networks consist of multiple domains. Each of these domains serves a specific area e.g. a distribution network or location like a secondary substation. It has to support individual requirements driven by the applications it has to serve. A Smart Grid communication solution needs to map those requirements to suitable communication technologies per domain and combine them to an integrated end-to-end communication network. The communication network architecture, shown in Figure 1, is derived from and mapped to the Power Network it serves. This approach provides an architectural Smart Communication Networks for Smart Grids which is comprehensive, simple and easy to understand. The two border elements of these voltage levels are of major relevance for the Smart Grid communication domains as they have their own very specific requirements: Secondary Substation: Low voltage to medium voltage transformation point. Primary Substation: Medium voltage to high voltage transformation point.

The following communication network domains can be distinguished:

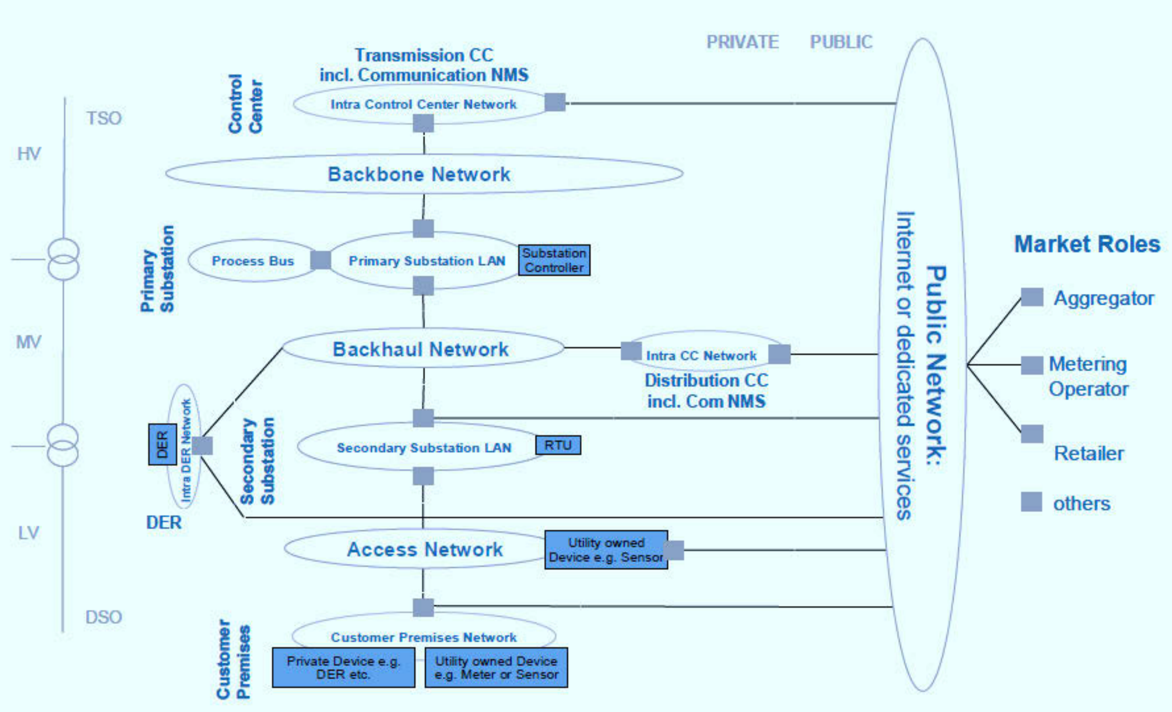


Fig. 1. Smart Grid Communication Domains

Backbone Network: Communication network which connects the Primary Substation LANs amongst each other and with regional control centers (often co-located) and central control centers.

Primary Substation LAN: A Primary Substation LAN is quite complex and requires an own communication infrastructure that distinguishes between a Process Bus and a Station Bus. It is mainly based on a Gigabit Ethernet infrastructure.

Backhaul Network: Communication network which connects the Secondary Substation LANs with each other and with a control center. This network domain might also connect to the respective Primary Substation LAN in case there is one stakeholder owning both, the medium voltage and high voltage power network.

Secondary Substation LAN: Network inside the secondary substation (today this network is quite trivial and consists of just one single Ethernet switch / IP router). The Secondary Substation LAN is implemented in US-Style regions more in a distributed manner whereas in Europe the Secondary Substation LAN is often located in an encapsulated enclosure.

Access Network: DSO-owned communication network which connects the customer premises or e.g. low-voltage sensors to a specific Secondary Substation.

Customer Premises LAN: In-building communication network whereas a customer is characterized by consumption and production of energy (Prosumer) and may be residential, public or industrial.

Intra DER Network: For medium-sized DERs like wind/solar parks a dedicated LAN is required for control, management and supervision purposes.

Intra-Control-Center Network: LAN within a DSO's or TSO's control center.

Public Network: Fixed or Mobile Network Operator owned communication network which offers different connectivity services either via dedicated services or via the Internet.

4 Smart Grid Applications

A huge variety of Smart Grid applications have been analyzed and assessed to identify their requirements on communication networks. They can be grouped into four application clusters: Metering Related Applications, Distribution Automation or Feeder Automation Applications, Management and Control of distribution network components incl. DERs and SCADA, and Auxiliary data transfer like Communication Network Management or CCTV etc. Based on the applications analyzed that are seen being implemented over the coming 5 years the requirements on the communication domains have been identified as mentioned in Table 1. Based on this understanding the communication network designer is able to select the right mix of communication technologies applicable for the specific environment.

The requirements driven by the applications that were discussed, the communication network domains and the grid topology lead to the recommended communication technologies. There two different types of the basic layout of a Medium Voltage Network need to be distinguished: Layouts that show long feeder lines with many transformers connected to serve few households or short feeders and transformers that serve hundreds of end customers. Typically the latter is what we call an "EU-Style scenario" while the former is classified as "US-Style scenario". This leads to the recommended use of Smart Grid Communication as shown in Table 2.

Table 1. REQUIREMENTS

BACKBONE NETWORKS AND INTRA-PRIMARY SUBSTATION NETWORKS	
Bandwidth	$0, 1 \ldots 1$ Gbps, up/down split irrelevant
Latency	5 ms
Availability	99,999% equal to 5 min downtime p.a.
Failure Convergence Time	seamless media redundancy needed
Battery Backup	Mandatory
BACKHAUL NETWORKS	
Bandwidth	1 to 2 Mbps (70% up, 30% down)
Latency	50 ms
Availability	99,99% equal to 50 min downtime p.a.
Failure Convergence Time	<1 s
Battery Backup	Mandatory
INTRA SECONDARY SUBSTATION NETWORKS	
Bandwidth	<500 kbps (70% up, 30% down)
Latency	50 ms
Availability	99,99% equal to 50 min downtime p.a.
Failure Convergence Time	<1 s
Battery Backup	Mandatory
ACCESS NETWORKS	
Bandwidth	1 kbps per residential user (70% up, 30% down)
Latency	<1 s
Availability	99% equal to 9 h downtime p.a.
Failure Convergence Time	<1 s
Battery Backup	Not required

Table 2. RECOMMENDED COMMUNICATION TECHNOLOGIES

NETWORK TIER	US STYLE	EU STYLE GRID LAYOUT
Backbone Network / Intra Primary Substation Network		Fiber Optic Networks
Backhaul Network	– Fiber Optic Networks – P2P Wi-Fi (long range) – WiMAX (private, specific utility design) – Public Cellular Broadband	– Fiber Optic Networks – WiMAX (private, specific utility design) – Public Cellular Broadband
Intra Secondary Substation	– Fiber Optic Network – Copper – RF Mesh / Wi-Fi Mesh	– Fiber Optic Network – Copper
Access Network	– RF Mesh – Wi-Fi Mesh – P2P Radio (long range) – Public Cellular – Public Fixed Network Internet Access (dedicated or open)	– Narrowband or Broadband PLC – Public Cellular – Public Fixed Network Internet Access (dedicated or open)

References

1. Electric Power Research Institute. Report to NIST on the Smart Grid Interop erability Standards Roadmap. (2009, June). [Online]. Available: `http://www.smartgrid.gov/document/report_nist_smart_grid_interoperability_standards_roadmap`
2. CEN-CENELEC-ETSI Smart Grid Coordination Group. Smart Grid Reference Architecture. (2012, November). [Online]. Available: `ftp://ftp.cen.eu/EN/EuropeanStandardization/HotTopics/SmartGrids/Reference_Architecture_final.pdf`
3. EU Commission Task Force for Smart Grids Expert Group 3. Roles and Responsibilities of Actors involved in the Smart Grids Deployment. [EG3 Deliverable]. (2011, April).

Author Index

University
of Bamberg
Press

At present, a comprehensive set of measurement, modeling, analysis, simulation, and performance evaluation techniques are employed to investigate complex networks. A direct transfer of the developed engineering methodologies to related analysis and design tasks in next-generation energy networks, energy-efficient systems and social networks is enabled by a common mathematical foundation. The International Workshop on „Demand Modeling and Quantitative Analysis of Future Generation Energy Networks and Energy-Efficient Systems“ (FGENET 2014) and the International Workshop on „Modeling, Analysis and Management of Social Networks and their Applications“ (SOCNET 2014) were held on March 19, 2014, at University of Bamberg in Germany as satellite symposia of the 17th International GI/ITG Conference on „Measurement, Modelling and Evaluation of Computing Systems“ and „Dependability and Fault-Tolerance“ (MMB & DFT 2014). They dealt with current research issues in next-generation energy networks, smart grid communication architectures, energy-efficient systems, social networks and social media. The Proceedings of MMB & DFT 2014 International Workshops summarizes the contributions of 3 invited talks and 13 reviewed papers and intends to stimulate the readers’ future research in these vital areas of modern information societies.

eISBN: 978-3-86309-209-2

9 783863 092092

www.uni-bamberg.de/ubp