



CORVINUS UNIVERSITY OF BUDAPEST

Doctoral School of Economics, Business and Informatics

Thesis Booklet

for the Ph.D. Dissertation by

Melinda Magyar

titled

**Analysis and Modeling of User Interactions and
Community Dynamics in Virtual Communities**

Supervisors:

Dr. Andrea Kő, Dr. Dávid Burka

BUDAPEST, 2025

Corvinus University of Budapest

Institute of Data Analytics and Information Systems

Thesis Booklet

for the Ph.D. Dissertation by

Melinda Magyar

titled

**Analysis and Modeling of User Interactions and
Community Dynamics in Virtual Communities**

Supervisors:

Dr. Andrea Kő, Dr. Dávid Burka

© Melinda Magyar

Contents

1. Research background and motivation 1

2. Methodology 4

3. Main findings of the dissertation 9

 3.1 Design science-based research framework and evaluation
 9

 3.2 Extensive literature review on two levels 9

 3.3 Unified data model for virtual communities (conceptual–
 logical–physical levels)..... 10

 3.5 Design and implementation of a reusable ETL process 11

 3.6 Demonstration with real-world data (Simple English
 Wikipedia)..... 11

 3.7 Reusable and well-documented dataset for research
 purposes 11

 3.8 Empirical insights derived from the Wikipedia community
 11

4. Main references..... 12

5. Publications of the author 15

1. Research background and motivation

The emergence of the internet and digital communication technologies has revolutionized social interactions, giving rise to virtual communities as new forms of social spaces. This phenomenon presents both challenges and opportunities for the social sciences, which have long studied the dynamics of communities and intercultural interactions. As these modern technologies become widespread, researchers equipped with traditional methodological tools must adapt to the changing environment and develop novel approaches to understand and analyze the complex social processes unfolding within data-rich virtual communities.

The evolution of virtual communities has opened new opportunities for computer science and data science as well, enabling the extraction of insights from data. Similar to social science research, computer and data sciences face challenges posed by the volume, heterogeneous structure, and sensitivity of the data. Integrating innovative technologies into research methodologies and handling data ethically are essential for gaining valuable insights into the diversity of community behavior and interactions in the digital age.

The traditional concept of human community implies the need for personal connection and acquaintance, and thus, implicitly, geographical proximity. This naturally limits the range of communities an individual can access and the extent of their possible participation. With the emergence of the internet and the accessibility of communities formed on online platforms, this constraint is now limited only by the amount of time a user is able to devote. Numerous communities compete for the time and attention of potential members, while various platforms compete to host these communities.

What makes virtual communities particularly compelling compared to traditional ones is that the behavior of members, their relationships, and the trajectory of the community can be observed from the position of a neutral observer, without the researcher influencing the community through their presence. Virtual communities operate in public view, and their communication is recorded in the form of data, predominantly in written form. This constitutes a vast repository of information, enabling an unprecedented level of insight into individual and collective behavior, the identity-shaping and norm-generating power of communities, and the life paths of individuals within the community – from joining to leaving the community.

To ensure the broad applicability of heterogeneous datasets, the use of generalized data models is a common practice in data modeling. This dissertation presents the development and application of a data model based on the structural analysis of virtual communities, designed to support comprehensive analysis and understanding of such communities. This approach enables researchers and decision-makers to access diverse community data from various sources in a unified, easily interpretable, and processable format.

The research places particular emphasis on user behavior and its changes throughout the community lifecycle, which is crucial to understanding why users leave virtual communities. In developing the model, I conduct a comprehensive examination of the dynamics of the individual, the community, and society – as well as their virtual counterparts – with special attention given to high-reach social platforms.

2. Methodology

The methodological foundation of the research is based on the design science approach, which combines elements of scientific inquiry and practical problem-solving. The aim of this methodology is to create artifacts – specifically, a unified data model and an ETL process description – that can effectively handle data from virtual communities and support their analysis. Design science allows for the continuous refinement of the problem space and the iterative development of proposed solutions, while rigorous evaluation mechanisms ensure the validity of the results. The research process closely followed the Hevner framework, which emphasizes a balance between relevance, rigor, and evaluation.

The first step of the process was the identification of the research problem and the definition of relevant objectives. This was accompanied by an extensive review of the literature and an examination of the data structures of various virtual community platforms. Key considerations in defining the problem space included the challenges of integrating heterogeneous data sources, accessibility constraints, and the alignment of differing analytical methodologies. The aim of the research is to develop a generalized data model capable of handling data generated by

various platforms and supporting the analysis of community interactions and behavioral patterns.

An iterative approach was applied throughout the research. Each iteration began with the refinement of a specific problem area, followed by the development of possible solutions, the creation of a prototype, and its evaluation. Based on the feedback received, the model was continuously refined, taking into account practical implementation aspects and the specific characteristics of various platforms. Over the course of the iterations, the structure of the data model evolved in multiple stages – from the conceptual model to the logical model, and finally to the physical data model.

Another key pillar of the methodological framework was the case study approach, through which the developed model and process were tested using real data from a specific community platform. The case study method proved particularly suitable for the in-depth examination of the complex phenomena of virtual communities, as it enabled the exploration of causal relationships and detailed contextual analysis. Both single-case and longitudinal case studies were employed: the former assessed the model's applicability on individual platforms, while the latter examined temporal changes in community dynamics.

The third methodological component of the research was the application of the CRISP-DM data mining process model. This framework structured the data processing and analysis steps of the research – from problem understanding through data preparation and modeling to evaluation and deployment. CRISP-DM provided a flexible structure that ensured data from various sources underwent a unified analytical process, while also allowing for continuous iterative feedback throughout the workflow.

A central element of the methodology was the creation of a unified data model. The model was built around three main entities – User, Context, and Content – interconnected through the interactions and relationships between them. It was capable of integrating various types of data, regardless of the specific characteristics of the source platform. During development, the strengths and weaknesses of existing data models were taken into account, as well as analytical requirements, with particular focus on examining community dynamics, user life cycles, and attrition.

The design of the data model was carried out in multiple stages. The process began with the identification of data requirements, which included the analysis and comparison of conceptual data

models from various platforms. This was followed by the first and second iterations of the unified conceptual data model, which gradually integrated elements from different systems. During the development of the logical data model, the refinement of entities and relationships, along with the consideration of normalization principles, ensured consistent and efficient data representation. The physical data model was aligned with the concrete implementation and the design of ETL (Extract-Transform-Load) processes.

In the evaluation phase, the functionality and practical utility of the model were demonstrated using a dataset spanning 21 years from Simple English Wikipedia. This platform provided an ideal test environment for the model, as it features complex community interactions and long-term dynamics. During the analysis, metrics were applied to measure user activity, the evolution of the social network, and community stability.

The chosen methodological framework offers the advantage of ensuring both scientific rigor and practical applicability. The combined use of design science, case study, and CRISP-DM made it possible for the research findings to be valuable to both the academic community and practical users. The overall structure of the dissertation is presented in Figure 1.

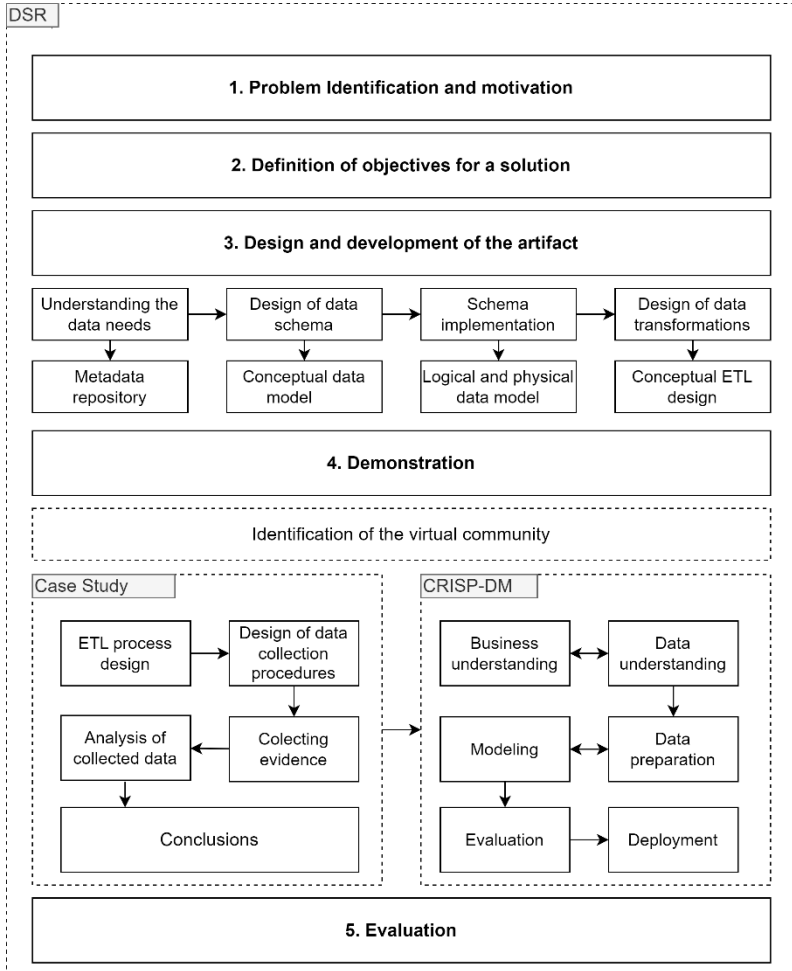


Figure 1: Methodological framework and structure of the dissertation

3. Main findings of the dissertation

The main scientific contributions of the research can be summarized in the following points.

3.1 Design science-based research framework and evaluation

The research covers the full cycle of DSR. Leveraging the flexibility of the framework, the demonstration of the results is supported by a custom methodology built on a metadata repository design standard, as well as case study and CRISP-DM-based approaches.

3.2 Extensive literature review on two levels

Grounded in both social science and data science, the research builds on a dual theoretical foundation to present the theoretical background of virtual communities, including role and lifecycle models (e.g., Rheingold, Plant, Golder & Donath, Preece & Shneiderman, Iriberry & Leroy), and integrates these into its own research objectives. It provides a detailed overview of the challenges and opportunities related to virtual communities as data sources.

A set of 20 publications focusing on virtual communities is analyzed, selected to maximize variation; the results are

structured by methods and metrics, including time-based, content-based, and network-based approaches.

3.3 Unified data model for virtual communities (conceptual–logical–physical levels)

By constructing and metamodeling the conceptual data models of five different types of platforms (social media, news feed, forum, project management, version control), the research established a foundation for comparative analysis across distinct community environments.

Through the generalization of data structures, an interoperable, platform-independent schema was developed that captures the key entities and relationship structures of diverse community systems.

3.4 EAV-based attribute management and temporal storage schema

The flexible structure, which also supports the time dimension, enables the management of heterogeneous community attributes (e.g., changes in user status) even within a relational environment.

3.5 Design and implementation of a reusable ETL process

Python-based preprocessing and a loading process built on Microsoft SQL Server.

The ETL steps cover data extraction, transformation (including cleaning, type handling, and missing data imputation), and loading.

3.6 Demonstration with real-world data (Simple English Wikipedia)

Demonstration of the model’s applicability to structurally diverse virtual communities, proving that the generalized model can be successfully applied across varying platforms.

The demonstration involved exploratory data analyses on editorial activity, user roles, context shifts, and life trajectories.

3.7 Reusable and well-documented dataset for research purposes

The created dataset is accessible, documented, and reusable for other research projects.

The documentation supports both reproducibility and repeatability.

3.8 Empirical insights derived from the Wikipedia community

The dataset generated during the demonstration was used for real research based on nearly two decades of events within an

existing community. The study shows that editing activity follows a highly skewed distribution (“long tail”), meaning that user life trajectories within the community tend to concentrate around an active core spanning multiple context types.

4. Main references

- Bégin, D., Devillers, R., & Roche, S. (2018). The life cycle of contributors in collaborative online communities - the case of OpenStreetMap. *International Journal of Geographical Information Science*, 32(8), 1611–1630. <https://doi.org/10.1080/13658816.2018.1458312>
- Bergstrom, K., & Poor, N. (2022). Signaling the Intent to Change Online Communities: A Case From a Reddit Gaming Community. *Social Media + Society*, 8(2), 1–10. <https://doi.org/10.1177/20563051221096817>
- Beuving, J. J. (2020). Ethnography’s future in the big data era. *Information, Communication & Society*, 23(11), 1625–1639. <https://doi.org/10.1080/1369118X.2019.1602664>
- Bruns, A. (2019). After the ‘APIcalypse’: social media platforms and their fight against critical scholarly research. *Information Communication and Society*, 22(11), 1544–1566. <https://doi.org/10.1080/1369118X.2019.1637447>
- Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, 1(1), 1–6. <https://doi.org/10.1177/2053951714540280>
- Davidson, B. I., Jones, S. L., Joinson, A. N., & Hinds, J. (2019). The evolution of online ideological communities. *PLOS ONE*, 14(5), 1–25. <https://doi.org/10.1371/journal.pone.0216932>

- Dessewffy, T., & Láng, L. (2015). Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*, 93(3–4), 157–170. http://www.replika.hu/system/files/archivum/92-93_11_dessewffy_lang.pdf
- Halford, S., & Pope, C. (2014). Big Data : Methodological Challenges and Approaches for Sociological Analysis. *Sociology*, 48(4), 663–681. <https://doi.org/10.1177/0038038513511561>
- Halford, S., & Savage, M. (2017). Speaking Sociologically with Big Data : Symphonic Social Science and the Future for Big Data Research. *Sociology*, 51(6), 1132–1148. <https://doi.org/10.1177/0038038517698639>
- Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems*. Springer US. <https://doi.org/10.1007/978-1-4419-5653-8>
- International Organization for Standardization. (2015). *Information technology – Metadata registries (MDR) (ISO Standard No. 11179:2015)*. <https://web.archive.org/web/20160406042038/http://metadata-standards.org/11179/>
- Iriberry, A., & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Computing Surveys*, 41(2), 1–29. <https://doi.org/10.1145/1459352.1459356>
- Ishikawa, H. (2015). Social Big Data Mining. In *Social Big Data Mining*. CRC Press. <https://doi.org/10.1201/b18223>
- Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, 4(1), 1–19. <https://doi.org/10.1186/s40537-017-0063-x>
- Preece, J., & Shneiderman, B. (2011). The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction*, 3(1), 1–25.

- Rheingold, H. (1993). *The Virtual Community. Homesteading in the Electronic Frontier*. Addison-Wesley Publishing Company.
<https://archive.org/details/virtualcommunity00rhei>
- Rowe, M. (2016). Mining User Development Signals for Online Community Churner Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(3), 1–28.
<https://doi.org/10.1145/2798730>
- Safari, R. M., Rahmani, A. M., & Alizadeh, S. H. (2019). User behavior mining on social media: a systematic literature review. *Multimedia Tools and Applications*, 78(23), 33747–33804. <https://doi.org/10.1007/s11042-019-08046-6>
- Sun, L. H., & Fichman, P. (2020). The collective trolling lifecycle. *Journal of the Association for Information Science and Technology*, 71(7), 770–783.
<https://doi.org/10.1002/asi.24296>
- Turner, T. C., Smith, M. A., Fisher, D., & Welser, H. T. (2005). Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*, 10(4), 1–37. <https://doi.org/10.1111/j.1083-6101.2005.tb00270.x>
- Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining. In *Social Media Mining: An Introduction*. Cambridge University Press.
<https://doi.org/10.1017/CBO9781139088510>

5. Publications of the author

- Magyar, M. (2020). *Perzisztencia és lemorzsolódás - Online motivációs tényezők vizsgálata adatbányászati eszközökkel a Wikipédia szerkesztői közösségében* [Corvinus University of Budapest]. http://szd.lib.uni-corvinus.hu/13512/1/Magyar_Melinda.pdf
- Magyar, M., & Burka, D. (2025). Using issue tracking as a groupwork facilitator in education. *Infocommunications Journal*, 17(Special Issue), 42–52. <https://doi.org/10.36244/ICJ.2025.5.6>
- Magyar, M., Kovács, L., & Burka, D. (2021). Forecasting the Spread of the COVID-19 Pandemic Based on the Communication of Coronavirus Sceptics. *The 7th International Conference on Time Series and Forecasting*, 1–10. <https://doi.org/10.3390/engproc2021005035>