CHExNet: A 400-years Multilayer Network of Early Modern Collaboration at the Jagiellonian University

Luiz do Valle Miranda[1[0000−0003−1838−5693]], Karol Dobiczek[1[0009−0001−0906−2393]], and Grzegorz J. Nalepa[1[0000−0002−8182−4225]]

Department of Human-Centered Artificial Intelligence, Institute of Applied Computer Science, Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, prof. Stanisława Łojasiewicza 11, 30-348 Kraków, Poland
luiz.miranda@uj.edu.pl, karol.dobiczek@doctoral.uj.edu.pl, grzegorz.j.nalepa@uj.edu.pl

Abstract. This paper introduces CHExNet, a temporal, two-layer network dataset capturing interactions centered on the Jagiellonian University in Kraków over the period 1364–1850. Nodes represent individuals reconstructed from heterogeneous archival and bibliographic sources, and edges encode two complementary interaction layers aligned to a shared time axis: (i) educational/professional co-presence, inferred when two persons are recorded at the same institution within the same semester-length bin, and (ii) book-production collaboration, inferred from co-participation in bibliographic records. The creation of the network is based on an expert-validated authority file reconciling person identities between the Jagiellonian University Archives and the Jagiellonian Library, and custom scripts for the creation of a series of time-based adjacency matrices. We describe the aggregated and temporal properties of the resulting network and illustrate its analytical value with a usage scenario that models the diffusion of Polish-language book production as a discrete-time hazard process driven by multiplex exposure. CHExNet is released for reuse as a benchmark for temporal and multilayer network analysis, historically grounded diffusion modeling, and AI methods that require long observation windows and interpretable provenance.

Keywords: Temporal networks · Multilayer networks · Social contagion

1 Introduction

Network science and computational social science have made significant progress in modeling collaboration, diffusion, and collective behavior using complex networks. Temporal-network approaches make it possible to study how the timing and ordering of interactions shape diffusion processes [**?**]. Empirical studies further show the predictive power of network topology on real world events, for example, clustered structure reflects a facilitation of behavior spreading when reinforcement matters, and that large-scale online cascades can exhibit systematically different spread patterns for true vs. false information [**?**]. At last, network

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

models can contain several layers, enabling analysis of how processes like diffusion interact across contexts rather than within a single aggregated graph [?].

The overall success of network-based approaches for studying social systems is tightly linked to the quantitative and qualitative availability of curated empirical datasets, i.e., networks constructed from observed or recorded data about real entities and relationships. A large set of reusable datasets have been released tracking the records of contemporary phenomena as complex networks. [?] presents a data set that tracks user behavior on the Bluesky social media platforms, including user-to-user and user-to-community interactions. Moreover, [?] shares and discusses face-to-face proximity networks collected with wearable sensors in households. Finally, [?] contains a large online-network corpora covering social, communication, citation, collaboration, and temporal networks.

While complex networks reflecting contemporary practices have the advantage of being large-scale, digitally recorded, and often time-stamped at fine resolution—making them attractive for measurement-driven modeling and benchmarking—one of their main limitations is the temporal span: many cover only a few years or decades, which prevents the study of slow, structural changes across centuries. One way to study such long-term dynamics is to work with historical social networks, which can extend the observation window substantially at the cost of noisier identifiers and sparser coverage. Examples include large-scale early modern social-network reconstruction projects such as Six Degrees of Francis Bacon [?], and historical communication/correspondence datasets such as Mapping the Republic of Letters, letter-based networks in early modern Europe [?]. A main limitation of historical social networks is that they are reconstructed from incomplete and heterogeneous records, so nodes and ties can be uncertain (name ambiguity), systematically missing (archival survival and social bias), and unevenly time-resolved, which can distort inferred structure and dynamics.

This paper introduces a temporal multilayer historical network spanning 1364–1850 centered around students and workers connected with the Jagiellonian University in Kraków within this period. In the created network, nodes represent individuals and edges capture in separate layers (i) co-presence at the same institution within a given period and (ii) book-production collaboration within the same period. Both layers are modeled as undirected, unweighted person–person graphs normalized to a common temporal axis. Section 2 we present the creation process for the network and information on its availability and additional content. Section 3 presents several properties of the network.

The resulting dataset is aimed at supporting research on temporal/multilayer network analysis, social simulation, and AI methods, thus Section 4 presents preliminary results of using CHExNet to answering two research questions. The paper ends with a summary of the work here presented, some of its limitations and provides directions for future work.

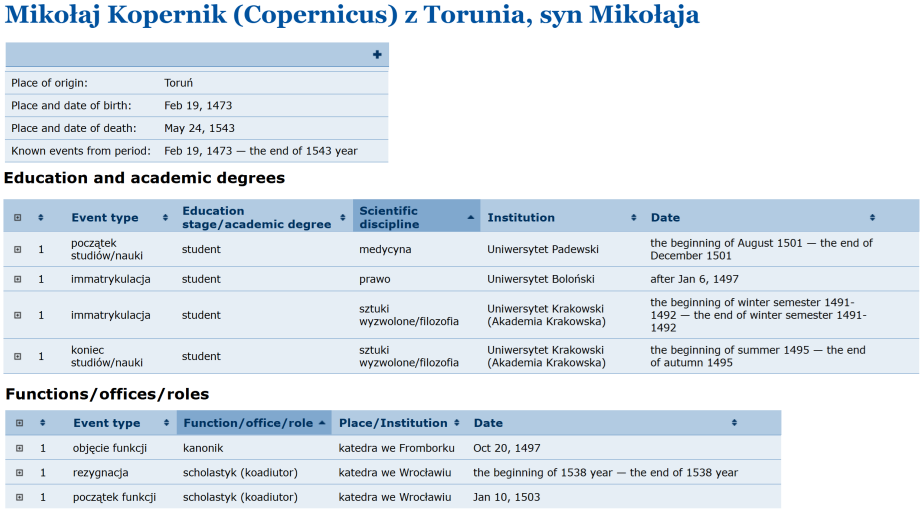
2 The Creation of CHExNet

2.1 Data Sources

The Jagiellonian University (JU), Poland's oldest university and among the world's oldest continuously operating institutions, houses a diverse collection

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

of cultural heritage artifacts connected to its more than 600 years history. Data concerning JU heritage resources are currently being stored and analyzed separately in different units, including the Archive (AUJ), and the Library (BJ). The ongoing CHExRISH[1] project at the JU has as one of its goals the unification of the data across the above-mentioned units. One of the consequences of such unification is the possibility of creating a multi-layered network of interaction between individuals connected with the university.



Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja

Place of origin:	Toruń
Place and date of birth:	Feb 19, 1473
Place and date of death:	May 24, 1543
Known events from period:	Feb 19, 1473 — the end of 1543 year

Education and academic degrees

	Event type	Education stage/academic degree	Scientific discipline	Institution	Date
1	początek studiów/nauki	student	medycyna	Uniwersytet Padewski	the beginning of August 1501 — the end of December 1501
1	immatrykulacja	student	prawo	Uniwersytet Boloński	after Jan 6, 1497
1	immatrykulacja	student	sztuki wyzwolone/filozofia	Uniwersytet Krakowski (Akademia Krakowska)	the beginning of winter semester 1491-1492 — the end of winter semester 1491-1492
1	koniec studiów/nauki	student	sztuki wyzwolone/filozofia	Uniwersytet Krakowski (Akademia Krakowska)	the beginning of summer 1495 — the end of autumn 1495

Functions/offices/roles

	Event type	Function/office/role	Place/Institution	Date
1	objęcie funkcji	kanonik	katedra we Fromborku	Oct 20, 1497
1	rezygnacja	scholastyk (koadiutor)	katedra we Wrocławiu	the beginning of 1538 year — the end of 1538 year
1	początek funkcji	scholastyk (koadiutor)	katedra we Wrocławiu	Jan 10, 1503

Fig. 1: Excerpt of Nicolaus Copernicus' record from CAC.

The main source of information for the creation of the network is Corpus Academicum Cracoviense (CAC)[2], an electronic database with around 67,000 records on students and graduates of the University of Kraków during the period 1364–1780[3]. CAC contains information on academic, occupational and personal events in a given person's life. Fig. 1 partially reproduces the CAC record for Nicolaus Copernicus, showing Copernicus's birth and death date, some institutions where he has worked, and places where he studied. CAC is transformed in a co-ocurrence network given the presence of an event at the same institution at the same time between two (or more) people, constituting layer 1 of CHExNet.

Layer 2 of CHExNet is created from the records contained in the ALMA system of BJ. ALMA is an integrated system for the management of around 9 million bibliographic records. ALMA data is structured according to the MARC21 format, thus containing information, among others, about authorship (e.g., authors, editors, translators, illustrators), publication facts (place, publisher, and date), language, and subject/genre metadata. Fig. 2 contains an excerpt of the bibliographic record for Copernicus' "De revolutionibus", where one can see the MARC21 fields and their respective values.

[1] For more information on the CHExRISH project and its use of Knowledge Graph technologies, see [?] and `https://chexrish.id.uj.edu.pl/`

[2] `https://cac.historia.uj.edu.pl/`

[3] See [?] for more details on the history of CAC and the provenance of its data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

MARC21 record (compact view)

001	9910885151605606
005	20240920061626.0
008	051121s19961520pl ||||gs|||||00| ||lat c
041	**a** lat
100	**a** Kopernik, Mikołaj **d** (1473–1543)
245	**a** De revolutionibus orbium coelestium **b** autograf : około 1520–1541 / **c** Mikołaj Kopernik.
700	**a** Zwiercan, Marian **d** (1932–). **t** Dzieje autografu *De revolutionibus* Mikołaja Kopernika

Fig. 2: Excerpt of selected MARC21 fields for Copernicus' "De revolutionibus".

A significant challenge for the creation of the multilayered network is the match between the two different bases. As visible from the name of Copernicus, in CAC the header is "Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja", and in ALMA it is "Kopernik, Mikołaj (1473–1543)". There was no authority file available for unifying these records. As described in [?], we have used a combination of string similarity metrics to train a random forest model to find possible matches between headers. The matches were further manually validated by domain experts from AUJ and BJ. A final authority file containing 824 verified matched entities was composed and was the basis for the creation of the network.

Having the authority file, we were able to retrieve the information from CAC and from ALMA related to the 824 individuals, summing up a total of 8658 bibliographical records in which at least one matched individual was an author or contributor and 48293 archival records about work or study activities of a matched individual. Fig. 3 presents the pipeline for the construction of CHExNet given the authority file and the CAC and ALMA bases. Custom python scripts were used to extract the data from the bases and merging them into CHExNet.

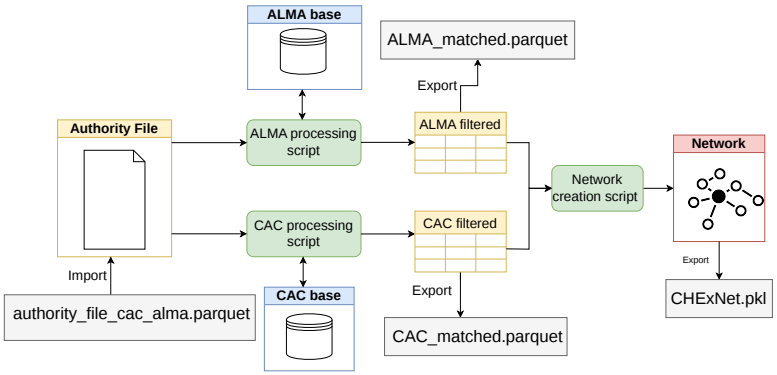


Fig. 3: Pipeline for the creation of CHExNet and the intermediary files.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

2.2 Network Modeling and Construction

We model *CHExNet* as a temporal multilayer network $G = (V, \mathcal{L}, \mathcal{T}, E)$, where V is the set of individuals, $\mathcal{L} = \{1, 2\}$ is the set of layers, $\mathcal{T}$ is a discretized time axis, and E is the set of time-resolved edges. The node set V is the set of collected individuals from both ALMA and CAC involved in the bibliographic and archival events[4] (currently $|V| = 3688$). Time is discretized into regular six-month bins spanning 1364–1850; we denote a bin by $t \in \mathcal{T}$ (e.g., $t = (y, \mathrm{H1})$ and $t = (y, \mathrm{H2})$ for the first/second half of year y).

Each layer is represented, for every time bin t, as an undirected and unweighted person–person graph $G_t^{[\ell]} = (V, E_t^{[\ell]})$. We interpret edges as binary indicators of a relationship being observed within the given time bin; if multiple records imply the same tie in the same bin, the edge is stored once (no weights). Self-loops are excluded, and edges are treated as undirected because both co-presence and co-contribution are symmetric relations in our operationalization.

Layer 1 captures educational and professional co-presence derived from CAC events. For each time bin t and each institution s, we consider the set of persons $P_{s,t} \subseteq V$ who have an event placing them at institution s during t. We then add an edge $(u, v) \in E_t^{[1]}$ for every unordered pair of distinct individuals $u, v \in P_{s,t}$, i.e., if they are recorded at the same institution in the same semester-length period. This construction yields a temporal sequence of co-occurrence networks representing opportunities for interaction or shared institutional context.

Layer 2 captures collaboration derived from the ALMA system. For each bibliographic record b with a publication date mapped to time bin t, we extract the associated contributors and map them to person identifiers when possible. For each time bin t, we connect all unordered pairs of distinct individuals who co-occur as authors or contributors on the same record, yielding edges $(u, v) \in E_t^{[2]}$.

2.3 Dataset Availability

The complete CHExNet dataset is publicly available via Zenodo [**?**]. The repository provides the two-layered network in a single *CHExNet.pkl* file, containing for each layer the adjacency matrices for all time bins. In the repository, the user can also find a brief documentation describing the file structure, temporal discretization, and basic usage. This representation supports standard temporal-network analyses (time slicing and aggregation) as well as multilayer methods that rely on aligned node sets and synchronized time bins.

To support reuse beyond purely graph-structural analysis, the release also includes two metadata tables: one containing publication-level information (*ALMA_matched.parquet*) and another describing individuals and a set of their associated events (*CAC_matched.parquet*). These metadata enable enrichment of network analyses (filtering, stratification, and attribute-aware modeling) and facilitate replication and extension of the usage scenario presented in Section 4.

For computational reproducibility, we additionally provide a public code repository containing the Python notebooks used to generate the descriptive

[4] There are individuals in V that are not in the authority file, since not every person in ALMA was matched.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

statistics, plots, tables, and the usage scenario reported in this paper. These notebooks serve as an assistance for understanding how to use the network[5].

3 Descriptive Statistics of CHExNet

3.1 Aggregated structure

Table 1 summarizes network-level metrics computed after aggregating all observed edges over the full time span, for the two layers (AUJ and BJ) and their union (Full). In terms of coverage, AUJ is much smaller (765 nodes) than BJ (3588 nodes), and Full adds only a limited number of additional nodes beyond BJ (3688), indicating that most actors in the union are already present in BJ.

Metric	**AUJ**	**BJ**	**Full**
Size (nodes)	765	3588	3688
Number of components	1	2	1
LCC size	765	3586	3688
LCC diameter	7	5	5
LCC edge density	0.086	0.070	0.068
LCC average path length	2.717	2.058	2.067
LCC average degree	66.241	251.494	252.890
LCC degree assortativity	-0.050	-0.011	-0.009
LCC clustering coefficient	0.769	0.768	0.752

Table 1: Aggregated network statistics for the two layers and their union.

All three aggregated graphs are (almost) fully connected. AUJ and Full each form a single connected component, while BJ splits into two components; however, BJ's largest connected component contains essentially all of its nodes (3586 out of 3588). In other words, when edges are treated as "ever observed" ties, the dataset exhibits a strong connected backbone with only a negligible disconnected remainder, and the union removes yields a single connected graph.

Distances in the largest connected component are short for all network versions (diameter 5–7; average path length approximately 2–3), implying high reachability: most pairs of individuals can be connected through only a few intermediaries. These short distances coexist with high clustering (0.75–0.77), indicating triadic closure and tightly interlinked local neighborhoods (e.g., cohorts or institutional clusters), while maintaining a compact global structure.

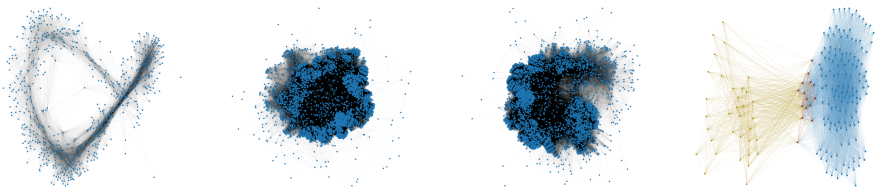
The layer-based networks differ more clearly in their level of concentration and local density. AUJ's Largest Connected Component (LCC) is denser than BJ and Full by edge density (0.086 vs. 0.070 and 0.068), yet it also has much lower average degree (66 vs. >250), reflecting that density and degree respond differently to network size and that the three graphs operate at different scales.

Figure 4 presents a visualization of the two layers separately and of their union as well as a temporal "slice" of the union between the years 1600 and 1605. Consistent with these aggregate statistics, the layouts in Figure 4 highlight a dense, visually saturated core in BJ and Full (a "hairball" effect caused by

[5] `https://github.com/luizdovalle2/CHExNET_Analysis`.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

heavy edge overlap), while AUJ appears as a smaller, more compact structure. The figure also presents a temporal slice of the network (1600-1605), showing at a smaller scale the interaction between the AUJ and BJ layers (colored yellow and blue, respectively), and their shared nodes and edges (colored red). This visualization shows the multiplex nature of the network, and thus how the two layers complement each other.



(a) AUJ network (b) BJ network (c) Full network (d) Temporal slice

Fig. 4: Visualization of the two layers, their union, and a temporal slice.

Degree assortativity in the LCC is slightly negative in all three layer-based networks (AUJ: -0.050, BJ: -0.011, Full: -0.010), indicating weakly disassortative mixing: highly connected nodes do not preferentially connect to other highly connected nodes, but the effect is modest. Importantly, the Full network preserves the short-distance/high-clustering regime while ensuring complete connectedness and maximal coverage, making it a natural baseline for subsequent analyses that require a single connected backbone.

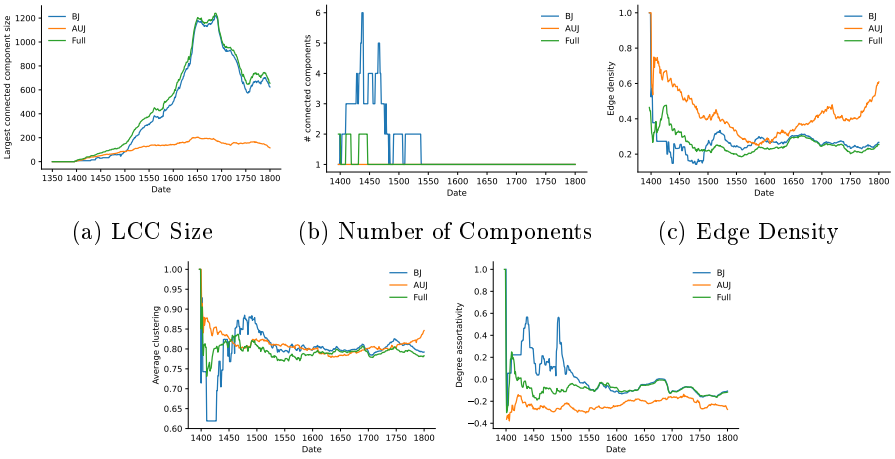
3.2 Temporal characteristics

To describe long-run change, we construct time-dependent graphs using a fixed persistence rule: once an edge appears, it remains active for 50 years.[6] At each time point we compute (i) the size of the LCC, (ii) the number of connected components, (iii) edge density, (iv) the average clustering coefficient, and (v) degree assortativity. Figure 5 presents charts for visualization of the above-mentioned metrics. We report these values only up to 1800; beyond this point the underlying network starts becoming very sparse, so the resulting statistics become unstable and harder to interpret reliably.

The LCC trajectory captures global integration under the rolling window: the Full network sustains the largest connected backbone across time, indicating that combining layers repeatedly adds bridging ties that keep more nodes mutually reachable. This interpretation is reinforced by the number of connected components: Full is connected (one component) for most snapshots, whereas BJ is fragmented into multiple components in the early period before consolidating. In other words, the union does not merely increase the size of the giant component; it often collapses a fragmented structure into a single connected system.

[6] This persistence window is a methodological choice that trades temporal resolution for interpretability: without persistence, the snapshots are extremely sparse and many structural metrics fluctuate strongly. The 50-year window provides a coarse approximation of medium-term institutional and collaborative "memory".

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4



(a) LCC Size (b) Number of Components (c) Edge Density

(d) Clustering Coefficient (e) Degree Assortativity

Fig. 5: Metric values of the temporal networks.

Edge density highlights a different division of roles. AUJ is the densest for most of the period, with a high density in the early 1400s, a decline into the 16th century, and a renewed rise after ∼1600 that accelerates toward 1800. In contrast, BJ is relatively sparse and more stable, with only slight mid-period increases. The Full density typically lies below the two and tracks BJ more closely, suggesting that the union's overall volume is constrained by the sparser BJ even when AUJ is internally dense; thus AUJ's contribution is not simply "more edges," but edges that connect otherwise separate portions of the graph.

Clustering coefficients in Full are generally comparable to, or slightly lower than, the single-layer curves, consistent with union edges acting more as cross-group bridges than as triangle-forming ties. Degree assortativity provides a complementary view of this mixing structure: AUJ remains persistently disassortative (high-degree nodes tending to connect to low-degree nodes), while BJ transitions from early assortative mixing to values closer to zero or mildly negative later on; Full is likewise mildly disassortative for much of the period. Taken together, these temporal signatures suggest complementarity between layers: AUJ concentrates dense within-layer connectivity and hub–periphery linking, BJ provides a sparser but stabilizing backbone, and their combination most consistently yields global integration (largest LCC and, frequently, a single connected component) without inflating local closure (clustering).

4 Usage Scenario: Social Contagion of Polish Writing

This section illustrates one use-case of CHExNet for studying a diffusion process in a historically grounded setting: the emergence and spread of the practices of producing books in Polish language[7] among individuals present in CHExNet in

[7] In preference to Latin

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

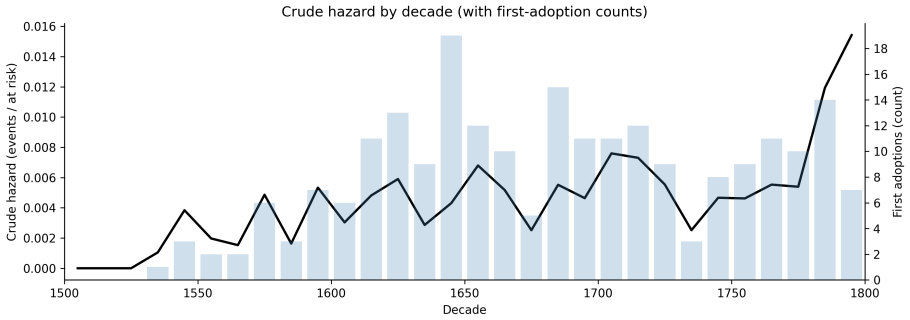


Fig. 6: Crude hazard of first Polish-language book production by decade (adoption events divided by the number at risk at the start of the decade). Bars show the number of first adoptions per decade.

the period 16–18 century. In summary, we want to present preliminary results of how CHExNet can be used to answer the following research questions:

> **RQ1**: Does exposure to prior Polish-language book production in a person's multiplex network increase the probability that the person begins producing Polish-language books?

Contagion hypothesis. Conditional on baseline time effects, an individual's hazard of *first* Polish-language book production increases with exposure to previously adopting neighbors in (a) the educational/professional co-presence layer and/or (b) the book-production collaboration layer. We further posit *cross-layer reinforcement*, i.e., the joint presence of exposures in both layers increases adoption risk more than either exposure alone.

> **RQ2**: Does a multiplex network structure adds predictive power for modelling social contagion in the case of Polish-language book production?

Multiplex hypothesis. A hierarchical multiplex exposure specification that includes both layer-specific exposure terms and a cross-layer interaction provides stronger predictive power than both layers collapsed into one exposure metric.

4.1 Data Operationalization and Discrete-Time Hazard Model

We model diffusion in discrete time using semestral bins. Individuals enter the risk set at their first observed activity in the underlying sources and remain at risk until first adoption; all person-bins after adoption are excluded to preserve a proper risk set. We define adoption as the first bin in which an individual appears as an author of a Polish-language book record, as indicated by bibliographic data. The outcome binary indicator $y_{i,t}$ takes value 1 if individual i adopts in bin t (first adoption) and 0 otherwise; with a total of 229 first-adoption events.

To provide a descriptive view of the diffusion dynamics, Fig. 6 reports the crude hazard by decade, defined as the number of first adoptions divided by the number of individuals at risk in that decade. The bars show first-adoption counts, while the line shows the corresponding hazard, allowing us to distinguish absolute event volume from relative adoption intensity.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

The figure indicates that adoption was rare in the early period, became more sustained from the late sixteenth and especially the seventeenth century onward, and fluctuated rather than rising monotonically. In the final decades of the sample, the hazard increases sharply, suggesting the highest adoption propensity at the end of the observation window. This descriptive pattern supports the inclusion of temporal controls in the hazard models and provides context for interpreting the diffusion results.

Row	% $x = 0$	% $x = 1$	% $x \geq 2$	% exposed
Layer 1	83.80	4.21	11.99	16.20
Layer 2	74.47	21.06	4.48	25.53
Either layer ($exp^{[1]}$=1 $\vee$ $exp^{[2]}$=1)				40.75
Both layers ($exp^{[1]}$=1 $\wedge$ $exp^{[2]}$=1)				0.99

Table 2: Exposure-count sparsity and any-exposure prevalence by layer; the last two rows report exposure in either layer ($\vee$) and simultaneous exposure in both layers (overlap, $\wedge$) for the person-year risk set.

For each individual i and bin t, we derive network exposures from the temporal graphs $G_t^{[1]}$ and $G_t^{[2]}$. Let $x_{i,t}^{[1]}$ and $x_{i,t}^{[2]}$ denote the number of neighbors of i at bin t that have already adopted by t in layer 1 and layer 2, respectively. Table 2 summarizes the resulting exposure distributions in the person-bin risk set and highlights the sparsity typical of historical reconstructions: in layer 1, $x_{i,t}^{[1]} = 0$ in 83.80% of person-bins (only 4.21% have exactly one exposed neighbor), while layer 2 remains sparse but with substantially more single-neighbor exposure (74.47% with $x_{i,t}^{[2]} = 0$ and 21.06% with $x_{i,t}^{[2]} = 1$). At the level of any exposure, 16.20% (layer 1) and 25.53% (layer 2) of person-bins have at least one exposed neighbor; exposure in either layer occurs in 40.75% of person-bins, whereas simultaneous exposure in both layers is rare (0.99%).

Given this sparsity, we use stable exposure encodings[8] based on any exposure: $\exp_{i,t}^{[1]} = \mathbb{I}(x_{i,t}^{[1]} > 0)$ and $\exp_{i,t}^{[2]} = \mathbb{I}(x_{i,t}^{[2]} > 0)$. We then estimate a discrete-time hazard model via logistic regression with baseline time effects captured by century fixed effects. Formally, for individual at risk i at time-bin t,

$$\begin{aligned} \operatorname{logit}\big(\Pr(y_{i,t} = 1 \mid y_{i,t-1} = 0)\big) = \alpha_{\text{century}(t)} \\ + \beta_1 \exp_{i,t}^{[1]} \\ + \beta_2 \exp_{i,t}^{[2]} \\ + \beta_3 \left(\exp^{[1]} \times \exp^{[2]}\right)_{i,t}. \end{aligned} \tag{1}$$

Here, $\alpha_{\text{century}(t)}$ absorbs macro-historical changes shared by all individuals in a century. We interpret $\beta_1 > 0$ and $\beta_2 > 0$ as evidence consistent with layer-specific diffusion, and $\beta_3 > 0$ as evidence consistent with multiplex reinforcement.

To assess whether network exposure improves fit beyond secular time trends, we compare Eq. 1 to a baseline model that includes only century fixed effects,

[8] For more information on threshold models and exposure operationalizations in diffusion/event-history settings, see [?].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

$$\operatorname{logit}\big(\Pr(y_{i,t}=1 \mid y_{i,t-1}=0)\big) = \alpha_{\text{century}(t)}. \tag{2}$$

We then estimate a small set of nested variants obtained by imposing restrictions of the form $\beta_k = 0$ (removing the corresponding exposure term), including models with only layer-1 exposure ($\beta_2 = \beta_3 = 0$), only layer-2 exposure ($\beta_1 = \beta_3 = 0$), and an additive two-layer model without reinforcement ($\beta_3 = 0$).

Finally, we also estimate a specification based on a *layer-collapsed* exposure metric, in which the two layers are combined, capturing whether an individual is exposed in *either* layer at time t. Concretely, we define $\exp_{i,t}^{[\text{or}]}$ as the logical union of the two any-exposure variables, i.e., $\exp_{i,t}^{[\text{or}]} = \exp_{i,t}^{[1]} \vee \exp_{i,t}^{[2]}$, and fit

$$\operatorname{logit}\big(\Pr(y_{i,t}=1 \mid y_{i,t-1}=0)\big) = \alpha_{\text{century}(t)} + \beta_{\text{or}} \exp_{i,t}^{[\text{or}]}. \tag{3}$$

This collapsed model provides a parsimonious baseline for RQ2, testing whether multiplex information can be summarized as a single "any exposure" signal, and it enables direct comparison to the hierarchical multiplex specification that retains layer-specific terms and cross-layer reinforcement.

Given the rarity of adoption events and the sparsity of exposure patterns, we estimate logistic hazard specifications using Firth's logistic regression to improve stability under rare events and potential (quasi-)separation [**?**]. We evaluate improvement over the baseline using likelihood-ratio tests for nested maximum-likelihood models and report effect sizes as odds ratios with 95% confidence intervals; Wald tests are used for individual (e.g., $\beta_3 = 0$) and joint hypotheses on exposure coefficients. The comparison results are reported in Table 3.

4.2 Results

Table 3 provides preliminary evidence on both research questions[9]. Relative to the baseline model with century fixed effects only (M0), adding exposure terms improves model fit in some specifications, indicating that network structure contains predictive signal beyond macro-historical time trends. Across single-layer models, exposure in the collaboration layer (M2) shows a clear association with adoption and improves fit relative to the century-only baseline (M0): OR($\exp^{[2]}$)=0.67 [0.48, 0.92], $p = 0.01$, and LR vs M0 = 10.16 (df=1), $p < 0.001$. Exposure in the co-presence layer alone (M1) is not statistically supported at conventional levels (OR($\exp^{[1]}$)=0.80 [0.54, 1.13], $p = 0.21$), although it still yields a modest improvement over baseline (LR vs M0 = 5.04, df=1, $p = 0.02$).

When both main effects are included additively (M3), both exposure terms are estimated below one ($\exp^{[1]}$: OR=0.70 [0.47, 1.00], $p = 0.05$; $\exp^{[2]}$: OR=0.62 [0.45, 0.86], $p < 0.001$), and the model improves over baseline (LR vs M0 = 17.43 (df=2), $p < 0.001$). Taken at face value, these results do not support a simple

[9] ORs are smaller than 1, which may appear counterintuitive if network exposure is interpreted as a simple monotonic driver of adoption. We do not read these coefficients as evidence that exposure "discourages" adoption. Rather, they likely reflect the fact that the measured exposure variables combine several sources of heterogeneity, including incomplete historical coverage, unequal layer size, and possible mismatch between the timing of observed ties and the timing of adoption.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

	M0	M1	M2	M3	M4	M5
Formula	$y \sim C(c)$	$+ \exp^{[1]}$	$+ \exp^{[2]}$	$+ \exp^{[1]} + \exp^{[2]}$	$+ \exp^{[or]}$	$+ \exp^{[1]} + \exp^{[2]} + \exp^{[1]} : \exp^{[2]}$
Log-likelihood	-1447.86	-1445.34	-1442.78	-1439.15	-1439.64	-1436.52
AIC	2901.72	2898.68	2893.56	2888.29	2887.28	2885.05
BIC	2928.26	2934.07	2928.95	2932.53	2922.66	2938.13
$\exp(\hat{\beta}_1)$ [95% CI]	–	0.80 [0.54,1.13]	–	0.70 [0.47,1.00]	–	0.63 [0.41,0.92]
p-value	–	0.21	–	0.05	–	0.02
$\exp(\hat{\beta}_2)$ [95% CI]	–	–	0.67 [0.48,0.92]	0.62 [0.45,0.86]	–	0.58 [0.41,0.80]
p-value	–	–	0.01	< 0.01	–	< 0.01
$\exp(\hat{\beta}_{\text{or}})$ [95% CI]	–	–	–	–	0.61 [0.46,0.80]	–
p-value	–	–	–	–	< 0.01	–
$\exp(\hat{\beta}_3)$ [95% CI]	–	–	–	–	–	3.99 [1.22,10.51]
p-value	–	–	–	–	–	0.03
LR vs M0	–	5.04	10.16	17.43	16.44	22.67
df diff	0	1	1	2	1	3
p (LR)	–	0.02	0.00	0.00	0.00	0.00
ΔAIC vs M0	0.00	-3.04	-8.16	-13.43	-14.44	-16.67
ΔBIC vs M0	0.00	5.81	0.69	4.27	-5.60	9.87

Notes: ORs are exponentiated coefficients from the logit model; p-values for OR rows correspond to Wald tests of the underlying coefficients. Likelihood-ratio (LR) tests compare each model to the baseline (M0).

Table 3: Model comparison across specifications. $C(c)$ denotes century fixed effects.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

"more exposure increases adoption" pattern in either layer individually; rather, exposure indicators are associated with lower estimated adoption odds unless additional structure (interaction) is accounted for.

The full hierarchical multiplex model (M5) provides the best fit by AIC among the reported specifications (AIC=2885.05) and yields the strongest likelihood improvement over baseline (LR vs M0 = 22.67 (df=3), $p < 0.001$). Crucially, M5 identifies a positive and statistically supported cross-layer interaction (OR($\exp^{[1]} \times \exp^{[2]}$)=3.99 [1.22, 10.51], $p = 0.03$), consistent with multiplex reinforcement: joint exposure across layers is associated with adoption risk beyond what would be implied by the two main effects alone. As in the previous analysis, the main effects remain below one ($\exp^{[1]}$: OR=0.63 [0.41, 0.92], $p = 0.02$; $\exp^{[2]}$: OR=0.58 [0.41, 0.80], $p < 0.001$), implying that the effect of exposure in one layer cumulates with exposure in the other.

Model M4 operationalizes a collapsed representation using a single exposure indicator ($\exp^{[or]}$) that flags exposure in either layer. This collapsed model fits better than the baseline (AIC=2887.28; LR vs M0 = 16.44 (df=1), $p < 0.001$), indicating that a one-dimensional "any exposure" summary retains substantial predictive signal. However, the full multiplex specification (M5) further improves AIC (2885.05 vs 2887.28) and substantially increases the log-likelihood (llf: -1436.52 vs -1439.64), suggesting that explicitly modeling layer-specific effects and cross-layer interaction captures additional structure that is lost under layer collapse. At the same time, BIC favors the collapsed model over the full multiplex model (M4: 2922.66 vs M5: 2938.13), reflecting BIC's stronger penalty for additional parameters; thus, the evidence for multiplex added value depends on whether one prioritizes predictive fit (AIC / likelihood) or parsimony (BIC).

In addition to the results calculated using half-year-long bins, we perform an analysis using 12 and 24 month long contagion periods. The results show a higher overall log-likelihood for these models with -820.81 and -782.52 for the 12 and 24 month window models (M0) respectively, an unsurprising fact, given the wider time windows. In the case of 12-month bins, the expansion of the time bin weakened the effects of the network structure on the fit of the model, the p-values increasing to $p \geq 0.05$ for all coefficients, except the exposure terms β_2 and β_3 in the full multiplex model (M5). This trend was present in the 24-month models; however, there the effect layer 1 of the network is also strengthened with p-values lower than 0.01 (M1 and M3). This variant also showed the highest log-likelihood increase over baseline (LR vs M0 = 31.54, (df=3), $p < 0.001$).

Overall, network exposure improves predictive performance beyond century-level baseline effects, but single-layer models do not support a simple monotonic "more exposure $\Rightarrow$ higher adoption" mechanism. The collaboration layer is much larger than the co-presence layer and provides most of the exposure variation in the union network, which likely explains why the clearest single-layer gains appear there. The best-performing specification is the multiplex model with layer-specific exposures and their interaction, implying that co-presence adds non-redundant information and that institutional proximity and collaboration-based exposure can jointly reinforce adoption.

A collapsed "any exposure" indicator provides a competitive baseline; however, this parsimony may partly reflect the dominance of the collaboration layer,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

since the collapsed signal is likely driven by collaboration exposure for most person-bins. We therefore hypothesize that the added value of explicitly modeling multiplex structure would become more pronounced in settings (or restricted subgraphs) where the two layers are more balanced in coverage and activity, e.g., within periods or institutional subsets where co-presence ties are denser, or within cohorts that participate comparably in both archival and bibliographic records. These results motivate robustness checks that vary the sampling frame (e.g., restricting to nodes observed in both sources, or to time periods with comparable layer activity) and exposure definitions, to assess when multiplex reinforcement is most detectable and to separate genuine cross-layer effects from layer-size and measurement imbalances. Overall, this suggests a hypothesis that a multiplex solution can be more robust independently of layer-based properties.

5 Conclusion

We introduced CHExNet, a two-layer temporal network (1364–1850) linking individuals connected to the Jagiellonian University via (i) educational/professional co-presence and (ii) book-production collaboration. The dataset reconciles heterogeneous archival and bibliographic sources into a unified person-level representation on a common time axis and is released with metadata and reproducible code. Aggregated descriptives show a well-connected backbone with short paths and high clustering, while temporal summaries track long-run changes in connectivity and highlight distinct layer contributions to the union graph. As an application, we model diffusion in Polish-language book production using a discrete-time hazard framework, comparing single-layer, collapsed, and multiplex exposure specifications; the results suggest that cross-layer interactions capture structure not recoverable from a single "any exposure" signal.

It is important to acknowledge the dataset's limitations. CHExNet is reconstructed from incomplete and heterogeneous historical records; node coverage and observed ties are uneven across periods, social groups, and institutions. Therefore, the absence of an edge does not imply the absence of a real-world relationship. Moreover, the two layers differ substantially in size and activity, which can dominate collapsed representations and complicate direct comparisons.

This release enables several extensions. The dataset supports richer multiplex and temporal analyses, such as community evolution, role and hub dynamics, and tests of layer interdependence beyond simple exposure counts. It also provides a basis for comparative studies of long-term collaboration and cultural diffusion, including replication of diffusion models across other topics, genres, or institutional contexts. Finally, future work will focus on the multiplex hypothesis to clarify when multiplex reinforcement is detectable and to distinguish genuine cross-layer effects from artifacts of layer size and sparsity.

Acknowledgments. This publication was funded by a flagship project "CHExRISH: Cultural Heritage Exploration and Retrieval with Intelligent Systems at Jagiellonian University" under the Strategic Programme Excellence Initiative at Jagiellonian University. The research for this publication has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4

We would like to thank AUJ and BJ for sharing access to their databases and for help in matching personal identifiers.

For this work, the authors used GPT-5.2 in order to perform: grammar and spelling check, paraphrase and reword. After using these services, the authors reviewed and edited the content and take full responsibility for the publication's content.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Marcin Baster. Corpus academicum cracoviense: database of students and professors of the university of krakow (1364-1780). In *Universitätsstudium und Gesellschaft in Mitteleuropa vom 15. bis zum 18. Jahrhundert*, volume 5 of *Historia et Monumenta Universitatis Jagellonicae*, pages 265–276. Towarzystwo Naukowe Societas Vistulana, Kraków, 2017.
2. Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.
3. Lorenzo Dall'Amico, Jackie Kleynhans, Laetitia Gauvin, Michele Tizzoni, L. Ozella, et al. Estimating household contact matrices structure from easily collectable metadata. *PLOS ONE*, 19(3):e0296810, 2024.
4. Kyle Dase. Six degrees of francis bacon. *Early Modern Digital Review*, 5(3).
5. Luiz do Valle Miranda, Krzysztof Kutt, and Grzegorz J. Nalepa. Cidoc-crm and the first prototype of a semantic portal for the chexrish project. In *Proceedings of the Second International Workshop of Semantic Digital Humanities (SemDH 2025), co-located with ESWC 2025*, volume 4009 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2025. CC BY 4.0.
6. Luiz do Valle Miranda, Maciej Mozolewski, Krzysztof Kutt, and Grzegorz J. Nalepa. A wikidata-based workflow for entity reconciliation strategies evaluation: A study on early modern polish personal names. In *Proceedings of ESWC 2026*, 2026. Accepted for publication (to appear).
7. Luiz do Valle Miranda and Grzegorz J. Nalepa. Chexnet: A network of collaborators in early modern jagiellonian university, 2026. Dataset. DOI: `https://doi.org/10.5281/zenodo.18715362`.
8. Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. Historical research in a digital age: Reflections from the mapping the republic of letters project. *The American Historical Review*, 122(2):400–424.
9. Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.
10. Petter Holme and Jari Saramäki. *Temporal Networks as a Modeling Framework*, pages 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
11. Ujun Jeong, Bohan Jiang, Zhen Tan, H. Russell Bernard, and Huan Liu. Bluetempnet: A temporal multi-network dataset of social interactions in bluesky social, 2024.
12. Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 07 2014.
13. Jure Leskovec and Rok Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.
14. Jiacheng Wu, Forrest W. Crawford, David A. Kim, Derek Stafford, and Nicholas A. Christakis. Exposure, hazard, and survival analysis of diffusion on social networks. *Statistics in Medicine*, 37(17):2561–2585, 2018.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_4