UNIVERSITY OF BATH

MSc DATA SCIENCE PAPER

Technology News Network Narrative Analysis

Author: Plato Ng

Supervisor: Nello Cristianini

Submitted in partial fulfilment of the requirements
for the degree of MSc Data Science
University of Bath, 2025

Narrative Network Analysis of Technology News (1987-2007)

submitted by

Plato Ng

for the degree of MSc in Data Science

at the University of Bath

2025

COPYRIGHT

Declaration

This paper is submitted to the University of Bath in accordance with the requirements of the degree of MSc in [Course Title] in the Department of Computer Science. No portion of the work in this paper has been submitted in support of an application for any other degree or qualification in any institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

The analysis of news narratives has traditionally relied on manual coding, which is resource-intensive and difficult to scale. Recent advances in computational methods offer new opportunities for automating this process, allowing the study of narrative structures across large corpora [Franzosi, 1987, Sudhahar et al., 2011]. This paper[1] applies network-based approaches to a corpus of technology news articles from the *New York Times*, combining centrality metrics with entity bias formulations to examine how key actors are positioned within technology narratives. The 2011 subject/object bias method [Sudhahar et al., 2011] is compared with the 2013/2015 weighting framework [Sudhahar et al., 2013, 2015], and a strong Spearman correlation ($\rho = 0.8471$) confirms the robustness of these measures. The results show that corporations such as Microsoft and IBM consistently occupy central positions, while financial and regulatory terms highlight the dual economic and institutional framing of technology discourse. Subject/object bias analysis further distinguishes between entities portrayed as initiators of actions (e.g., corporations, entrepreneurs) and those cast as recipients (e.g., technological artefacts, infrastructures). The study contributes a scalable framework for automated narrative analysis, offering insights into historical technology narratives and their potential real-time applications.

[1] Ethical approval was obtained prior to research (project ID: 11943).

Acknowledgements

My heartfelt thanks go to my supervisor, Nello Cristianini, whose expertise and patient guidance transformed my ideas into reality, making this project a rewarding journey. To my friends and coursemates in Bath and Leeds, your support, laughter, and late-night study sessions were my lifeline, turning challenges into cherished memories. To my family, you made studying in beautiful Bath possible. Mama, your timely food packages were my salvation; Dad, your steadfast support kept me going; and my elder sister Yolanda, the laughter you brought was the spark that brightened even the toughest days.

Contents

1 Introduction

1.1 Background and Motivation

The period from 1987 to 2007 represents a pivotal era in technological innovation, marked by the rise of personal computing, the spread of the internet, and the commercialisation of digital technologies. Milestones during this period include Microsoft's release of successive versions of the Windows operating system, IBMs continued dominance in enterprise computing, and Apples reinvention of consumer electronics with products such as the iPod [Campbell-Kelly and Aspray, 2003, Moschovitis et al., 1999]. The late 1990s in particular saw the dot-com boom, when venture capital fuelled rapid growth of internet companies, followed by the dot-com crash of 2000–2001, which reshaped both markets and public perceptions of technology [Cassidy, 2002]. The expansion of e-commerce, mobile telephony, and broadband infrastructure further created new digital markets that directly influenced economic and social life [Castells, 2001].

Understanding how these transformations were narrated in the media is crucial, as journalism not only reflects but also shapes public perceptions of technological change. Media discourse plays a role in framing companies as innovators or monopolists, technologies as disruptive or risky, and government as either enabler or regulator of digital economies. Such framings contribute to agenda-setting and shape industry reputation, investment, and policy responses [McCombs, 2005, Agarwal et al., 2012a]. By analysing narratives at scale, it is possible to uncover how corporate actors (e.g., Microsoft, IBM, Apple), technologies (e.g., the internet, email, mobile phones), and institutions (e.g., government, regulators) were positioned in relation to one another over time.

Traditional approaches to narrative analysis in media studies have relied on manual coding and close reading, which provide deep insights but are limited by issues of scale, subjectivity, and reproducibility [Franzosi, 2010]. These methods struggle to keep pace with the growing volume of digital content. To address these limitations, recent work in computational narratology has proposed automated methods to extract and analyse subject-verb-object (SVO) triplets from text, which can then be assembled into semantic networks that represent relationships between entities [Sudhahar et al., 2011, 2015]. Such approaches enable the systematic study of narratives across large corpora, combining the interpretive depth of narrative theory with the scalability of computational methods. This provides a novel means of examining how technological developments were represented in the media during a critical period of digital transformation.

1.2 Problem Statement

While prior studies have demonstrated the feasibility of automating narrative network analysis (NNA), much of this work has concentrated on domains such as crime reporting or political communication, where narrative roles are clearly defined and socially salient [Sudhahar et al., 2011, 2013]. In contrast, the application of NNA to technology news remains underexplored, despite the central role of media discourse in shaping public understanding of digital transformations and corporate reputation. This gap represents a missed opportunity, as the period between 1987 and 2007 was marked by the emergence of global technology corporations, and the rise of new markets that fundamentally altered economic and social life [Castells, 2001, Cassidy, 2002].

A second gap lies in the comparative evaluation of different formulations of entity bias. The subject/object bias method introduced by Sudhahar et al. [Sudhahar et al., 2011] provides a syntactic perspective on agency, distinguishing between entities acting as initiators versus recipients of actions. By contrast, the relevance weighting approach developed in later work [Sudhahar et al., 2013, 2015] captures domain-specific salience by comparing entity frequencies in a domain corpus against a background corpus, without differentiating grammatical roles. While both methods have been applied successfully in isolation, their degree of consistency when applied to the same technology corpus lacks systematic evaluation.

This paper addresses these gaps by applying automated NNA methods to a large-scale corpus of technology news from the *New York Times* Annotated Corpus (1987–2007). It compares subject/object bias and relevance weighting formulations within the same dataset, evaluates their consistency, and demonstrates how these methods can be applied longitudinally to trace the changing prominence and roles of corporate, technological, and institutional actors in the media.

1.3 Research Objectives

This paper develops an automated framework for narrative network analysis applied to technology news in the *New York Times* Annotated Corpus. The main objectives are:

1. Preprocess and filter technology-tagged articles from the New York Times Annotated Corpus (1987–2007) for compatibility with modern NLP tools such as spaCy, Coreferee, and Stanza.

2. Extract Subject-Verb-Object (SVO) triplets using neural dependency parsing and coreference resolution, ensuring accurate identification of entities and actions.

3. Construct semantic graphs representing relationships between entities, applying centrality measures to identify influential actors and structural patterns in the technology discourse.

4. Compute entity bias scores using both the 2011 subject/object method and the 2013/2015 weighting approach, comparing their results and evaluating consistency.

5. Validate findings against historical technology events and assess the methodological robustness of automated narrative analysis for large corpora.

1.4 Structure Overview

The remainder of this paper is organised as follows. Section 2 reviews related work, beginning with foundations in quantitative and narrative network analysis, before turning to developments in NLP tools, entity bias formulations, and broader perspectives from media and technology studies. Section 3 details the methodology, including data sources, preprocessing, triplet extraction, entity normalisation, weighting, and the construction of directed, weighted narrative networks. It also outlines the validation and reliability checks undertaken to ensure robustness. Section 4 presents the results, combining centrality measures, community detection, and entity bias analysis, followed by a comparative evaluation of bias formulations. Section 5 provides a discussion of the findings, addressing key limitations, methodological contributions, and the broader implications for the study of technology narratives. Finally, Section 6 concludes by summarising the contributions of this paper and outlining directions for future research.

2 Related Work

2.1 Narrative Network Analysis in Technology

Narrative network analysis (NNA) extends the principles of quantitative narrative analysis (QNA) by extracting SVO triplets from text and mapping them into networks of entities and actions. In this framework, entities are represented as nodes and their interactions, captured through verbs, are represented as directed edges. This transformation enables the study of narratives not only as sequences of events but also as structured systems of relationships among actors [Franzosi, 2010].

Early applications of NNA demonstrated the value of this approach in domains such as crime and political reporting, where actors such as perpetrators, victims, politicians, and institutions occupy distinct narrative roles that can be systematically quantified [Sudhahar et al., 2011]. For example, NNA has been used to uncover how suspects, law enforcement, and judicial institutions are linked within crime narratives, or how political leaders and policies are framed in news discourse. By turning unstructured text into analysable networks, these studies were able to identify recurring patterns, influential actors, and the structural features of narrative framing.

More recent work has extended NNA to large-scale corpora, making it possible to track the prominence of actors and the evolution of narratives over time [Sudhahar et al., 2013, 2015]. This shift from small-scale qualitative studies to automated, quantitative methods has enabled researchers to analyse hundreds of thousands of articles, providing both breadth and historical depth. In the process, NNA has shown how media narratives can be studied as dynamic systems that evolve with broader social, political, and economic change.

Within the domain of technology, narrative approaches have begun to highlight the role of corporations, innovations, and institutions in shaping public discourse [Agarwal et al., 2012a]. For instance, companies such as Microsoft, IBM, and Apple are often framed as central actors, while governments and regulators appear as mediators of technological and market change. However, relatively few studies have explicitly applied NNA to technology news in a longitudinal manner. Most work has concentrated on short periods or specific case studies, leaving open questions about how narratives evolve during critical periods such as the commercialisation of the internet.

2.2 Existing NLP Tools and Techniques

The implementation of narrative network analysis depends critically on the accuracy of natural language processing (NLP) components such as dependency parsing and co-reference resolution. Early work in this area often relied on rule-based or symbolic parsers, including Minipar [Lin, 1998], which, while innovative for its time, suffered from low recall, limited domain coverage, and compatibility issues with modern text formats. Similarly, early versions of the Stanford Parser [Klein and Manning, 2003] were based on probabilistic context-free grammars, which struggled with long-distance dependencies and complex syntactic structures frequently found in journalistic writing. These limitations restricted the scalability and reliability of automated triplet extraction in large corpora.

Advances in neural network approaches have substantially improved both precision and scalability. SpaCy [Explosion AI, 2022], one of the most widely used industrial-strength NLP libraries, provides efficient tokenisation, part-of-speech (POS) tagging, and dependency parsing. When combined with the Coreferee plugin [Hudson, 2023], SpaCy achieves competitive performance in co-reference resolution, consistently outperforming earlier neuralcoref implementations [Clark and Manning, 2016]. This step is crucial for narrative analysis, as unresolved pronouns or nominal anaphora (e.g., "he," "the company") fragment entities across multiple nodes in the network, artificially lowering their centrality.

Another widely adopted tool is Stanza [Qi et al., 2020], a neural dependency parser developed by the Stanford NLP Group. Based on universal dependencies, Stanza achieves Labeled Attachment Scores (LAS) exceeding 90% on benchmark datasets, substantially outperforming rule-based systems such as Minipar [Lin, 1998]. In addition, Stanza's multilingual support and modular pipeline make it particularly suited for cross-domain and cross-linguistic studies, expanding the applicability of narrative network analysis beyond English-only corpora.

These advances collectively resolve key methodological bottlenecks in earlier work, allowing triplets to be extracted with high reliability across hundreds of thousands of articles and documents.

2.3 Entity Bias Analysis

Entity bias analysis provides a complementary perspective by quantifying how entities are positioned within narratives as either initiators or recipients of action. Sudhahar et al. [2011] introduced a subject/object bias measure that operationalises this distinction using frequency distributions. For each actor K, the subject bias is defined as:

$$\text{Subject Bias}(K) = \frac{f_K(\text{Subject, Domain})}{f_K(\text{Subject, Background}) + f_K(\text{Object, Background})}, \tag{1}$$

where f_K represents the frequency of actor K in subject or object position within either the domain corpus (e.g., crime or technology news) or a background corpus (e.g., top news). A high subject bias indicates that an entity is frequently portrayed as an initiator of actions, while a low or negative score suggests passivity, reflecting its tendency to appear as an object of actions.

Subsequent work refined this framework by proposing a relevance weighting measure that captures entity salience relative to a background corpus, without distinguishing between syntactic roles [Sudhahar et al., 2013, 2015]. This weight is calculated as:

$$w_K = \frac{f_K(\text{Domain})}{f_K(\text{Background})}, \tag{2}$$

where $f_K(\text{Domain})$ is the frequency of entity K in the target corpus and $f_K(\text{Background})$ is its frequency in the background corpus. Actors with higher weights are considered disproportionately represented in the domain discourse, regardless of whether they act as subjects or objects.

These two formulations capture different but complementary dimensions of narrative structure: the subject/object bias focuses on grammatical role and agency, while relevance weighting emphasizes domain-specific salience. Despite their conceptual differences, both aim to identify entities that play central roles in shaping narratives. However, few comparative evaluations have examined the degree of consistency between the two methods. This paper addresses this specific gap by applying both approaches to the same technology news corpus and measuring their convergence through correlation analysis.

2.4 Scalability and Computational Challenges

A persistent challenge in narrative network analysis lies in scalability. The analysis of large corpora requires efficient filtering, parsing, and network construction techniques. Sudhahar et al. [Sudhahar et al., 2015] demonstrated that automated pipelines could handle large datasets, but noted issues of processing time and memory consumption. Studies in computational social science likewise emphasize the importance of batch processing and parallelization for scaling narrative analysis frameworks [Earl et al., 2004]. This paper contributes to this line of work by implementing compute-efficient techniques, including multi-core batch parsing, frequency-based filtering, and domain-specific relevance weighting. These refinements ensure that narrative extraction can be applied to more than 500,000 technology-tagged articles in the *New York Times* Annotated Corpus, while maintaining methodological validity.

2.5 Media Narratives and Technology Studies

While computational approaches to narrative analysis focus on methods for extracting and quantifying patterns in text, it is equally important to acknowledge the broader role of media narratives in shaping public understanding of technology. The sociology of technology and media studies literatures have long emphasised that news discourse is not a neutral reflection of technological developments, but an interpretive framework that positions actors, innovations, and institutions within wider cultural and political contexts [Flichy, 1995, Wyatt, 2008].

Media narratives exert influence through mechanisms such as framing and agenda-setting, which determine not only which issues are covered but also how they are presented [McCombs, 2005]. For example, narratives of technological progress often frame corporations as innovators and governments as regulators, while economic reporting foregrounds metrics such as profit, market share, and growth [Boczkowski, 2004]. These discursive practices contribute to the construction of collective understandings of technology, legitimising some actors while marginalising others.

In the context of digital transformation, [Castells, 2001] argues that media play a dominant role in shaping the "network society," where flows of information define social and economic structures. Similarly, [Wyatt, 2008] highlights how narratives of inevitability, disruption, and progress influence both public attitudes and policy debates around technology. Such perspectives underscore why studying media representations of technology is essential for understanding not only the diffusion of innovations but also the social meanings attached to them.

Despite this, relatively few studies have combined these theoretical insights with automated methods such as narrative network analysis. Bridging media studies and computational narratology provides an opportunity to capture both the structural properties of narratives and their interpretive functions, making it possible to trace how corporations, innovations, and institutions were framed during critical periods such as the commercialisation of the internet and the dot-com bubble. This paper seeks to contribute to this emerging intersection by applying narrative network methods to a large-scale corpus of technology news, thereby linking computational approaches with broader debates in media and technology studies.

2.6 Summary and Research Gap

This chapter has reviewed the key strands of literature relevant to this study. To begin with, research on quantitative and narrative network analysis has demonstrated that SVO triplets can be systematically extracted from text and mapped into semantic networks, enabling the study of agency and relationships among actors [Franzosi, 2010, Sudhahar et al., 2011]. Applications in crime and political reporting have shown the value of these methods for revealing how narratives structure roles and events over time [Sudhahar et al., 2013, 2015].

Developments in natural language processing have addressed earlier methodological bottlenecks. Whereas rule-based systems such as Minipar struggled with coverage and accuracy [Lin, 1998], contemporary neural approaches such as Stanza and SpaCy now achieve state-of-the-art performance in dependency parsing and co-reference resolution, making large-scale triplet extraction both scalable and reliable [Qi et al., 2020, Hudson, 2023].

Also, media and technology studies emphasise that narratives are not neutral descriptions but interpretive frameworks that frame corporations, technologies, and institutions within broader cultural and political contexts [Castells, 2001, Wyatt, 2008]. While this literature presents the importance of studying how technology is represented in the media, relatively few studies have combined these theoretical insights with automated methods of narrative analysis.

Taken together, these strands of literature highlight a gap that this paper seeks to address. Specifically, while narrative network analysis has been successfully applied to domains such as crime and politics, its application to technology news remains limited. Moreover, different formulations of entity bias-subject/object bias [Sudhahar et al., 2011] versus relevance weighting [Sudhahar et al., 2013, 2015] have not been systematically compared within the same technology corpus. Finally, there has been little longitudinal analysis of how technology narratives evolve during key historical periods such as the commercialisation of the internet, the dot-com boom, and its aftermath.

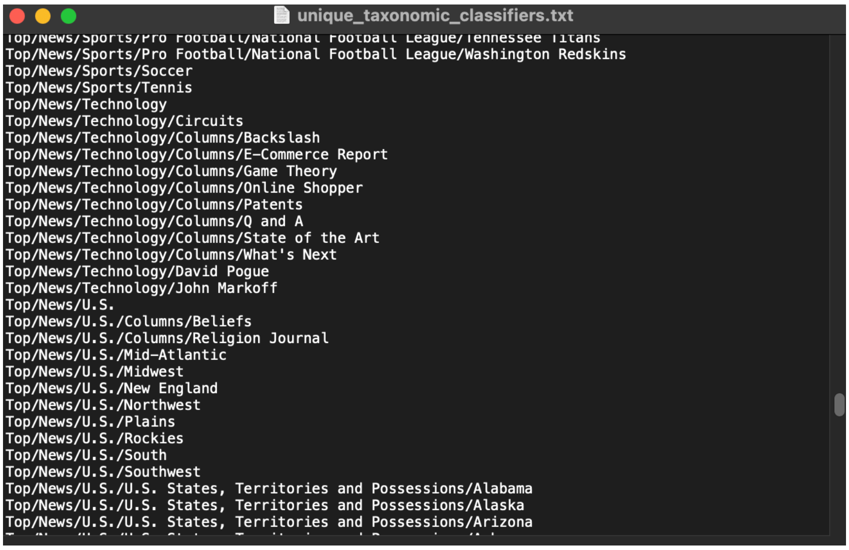
This paper addresses these gaps by applying automated narrative network analysis to two decades of technology reporting in the *New York Times* Annotated Corpus (1987-2007). By comparing alternative bias formulations and examining how corporate, technological, and institutional actors were positioned over time, the study contributes both methodological and empirical insights to the intersection of computational narratology and media studies.

3 Methodology

3.1 Data Source

The analysis in this paper is based on the *New York Times* Annotated Corpus (NYTAC), which contains over 1.8 million articles published between 1987 and 2007, each accompanied by detailed metadata such as publication date, section, and manually assigned descriptors [Sandhaus, 2008]. The dataset is widely used in computational linguistics and digital humanities because of its scale, historical depth, and consistent annotation standards.

For the purposes of this study, a subset of articles was selected using the taxonomy tags provided by the corpus. Specifically, articles annotated with `Top/News/Technology` or closely related descriptors were retained. Figure 1 provides an example of the hierarchical structure of taxonomic classifiers in the NYTAC, illustrating how articles can be filtered into domain-specific subsets. This filtering process resulted in approximately 500,000 technology-tagged articles, representing a focused corpus for examining narratives in the technology domain.



```
Top/News/Sports/Pro Football/National Football League/Tennessee Titans
Top/News/Sports/Pro Football/National Football League/Washington Redskins
Top/News/Sports/Soccer
Top/News/Sports/Tennis
Top/News/Technology
Top/News/Technology/Circuits
Top/News/Technology/Columns/Backslash
Top/News/Technology/Columns/E-Commerce Report
Top/News/Technology/Columns/Game Theory
Top/News/Technology/Columns/Online Shopper
Top/News/Technology/Columns/Patents
Top/News/Technology/Columns/Q and A
Top/News/Technology/Columns/State of the Art
Top/News/Technology/Columns/What's Next
Top/News/Technology/David Pogue
Top/News/Technology/John Markoff
Top/News/U.S.
Top/News/U.S./Columns/Beliefs
Top/News/U.S./Columns/Religion Journal
Top/News/U.S./Mid-Atlantic
Top/News/U.S./Midwest
Top/News/U.S./New England
Top/News/U.S./Northwest
Top/News/U.S./Plains
Top/News/U.S./Rockies
Top/News/U.S./South
Top/News/U.S./Southwest
Top/News/U.S./U.S. States, Territories and Possessions/Alabama
Top/News/U.S./U.S. States, Territories and Possessions/Alaska
Top/News/U.S./U.S. States, Territories and Possessions/Arizona
```

Figure 1: Snapshot of the taxonomic classifiers list from the NYT Annotated Corpus.

In line with [Sudhahar et al., 2011], the analysis also requires a background corpus to enable the computation of relative weights and bias scores.

For this purpose, the broader NYT Top News category was used, consisting of a large set of articles across diverse domains such as politics, economy, and society. This background corpus provides a reference distribution against which the prominence of technology-specific entities can be measured.

The choice of the NYT corpus is justified by three considerations. First, the *New York Times* is a globally recognised newspaper of record, offering comprehensive coverage of technology developments during the period of interest. Second, the twenty-year span of the dataset (1987–2007) coincides with key milestones in computing, the rise of the internet, and the dot-com bubble, making it an ideal source for longitudinal narrative analysis. Finally, the availability of detailed metadata facilitates reproducible filtering, temporal analysis, and integration with narrative network methods.

3.2 Design Pipeline and NLP Tools

The preprocessing pipeline prepared the raw *New York Times* Annotated Corpus for narrative network analysis by systematically transforming annotated XML articles into structured triplets. This process followed a multi-stage workflow, shown in Figure 2, combining text cleaning, co-reference resolution, and neural dependency parsing to ensure reliable extraction of entities and actions.

The key preprocessing steps were as follows:

- **Article Extraction and Filtering:** Articles were filtered by taxonomy tags (see Figure 1), retaining those classified under `Top/News/Technology`. Non-relevant items such as editorials, obituaries, and duplicates were excluded.
- **Co-reference and Anaphora Resolution:** To address pronoun ambiguity and repeated references, co-reference resolution was applied using the Coreferee plugin for spaCy [Hudson, 2023]. This ensured that terms such as "he," "it," or "the company" were consistently linked to their antecedents across sentences. Anaphora resolution improved entity continuity across longer spans of text.
- **Dependency Parsing:** Grammatical structures were parsed using Stanza [Qi et al., 2020], a neural dependency parser achieving over 90% LAS on English benchmarks. Compared with rule-based systems such as Minipar [Lin, 1998], Stanza offered greater accuracy and robustness for large-scale analysis.
- **Triplet Extraction and Storage:** From parsed sentences, Subject-Verb-Object triplets were extracted and stored in a structured JSON format. Each record included metadata (article ID, publication date, taxonomy tags) alongside resolved entities and their grammatical roles.
- **Preparation for Analysis:** Extracted triplets were filtered to retain only those containing key actors (e.g. corporations, technologies, institutions) and semantically meaningful verbs. Entities were normalised to avoid duplication. The resulting dataset was then ready for weighting, computation of bias measures, and network construction (Sections 3.3 and 3.4).

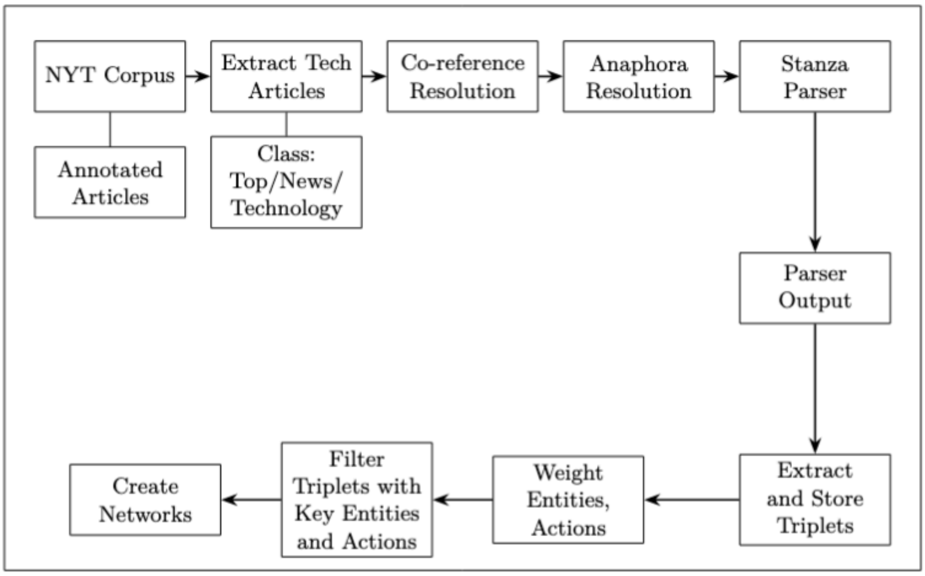


Figure 2: System pipeline for narrative network analysis, from annotated articles to structured triplets and network construction. Adapted and extended from [Sudhahar et al., 2011, 2013].

3.3 Triplet Extraction and Weighting

Following preprocessing, triplets were extracted to represent narrative events in structured form. These triplets capture relationships among entities (subjects and objects) mediated by actions (verbs), forming the building blocks of narrative networks. Prior work in quantitative narrative analysis and computational linguistics has demonstrated the value of SVO-based representations for modelling agency and event structure across large corpora [Franzosi, 2010, Poria et al., 2019]. By reducing complex syntax to relational units, triplets provide a scalable and reproducible method for mapping narratives into network form.

3.3.1 Triplet Extraction

From the dependency-parsed sentences produced in Section 3.2, SVO triplets were identified by linking the grammatical subject, the root verb, and the direct object of each clause. This approach draws on syntactic dependency structures, which have been shown to capture relational semantics more effectively than surface-level co-occurrence [de Marneffe et al., 2006, Manning et al., 2014]. Additional linguistic relations such as compound nouns (e.g., “New York Times”) and phrasal verbs (e.g., “set up”) were merged to ensure semantic coherence. In line with best practices in narrative extraction, triplets without a valid subject or object were discarded to minimise noise and avoid fragmentary edges in the resulting network [Sudhahar et al., 2011].

3.3.2 Entity and Verb Normalisation

To avoid duplication and fragmentation in the network, extracted entities were normalised through case standardisation, lemmatisation, and spelling disambiguation. For instance, “IBM” and “I.B.M.” were merged into a single entity node, while verb forms such as “develops” and “developed” were reduced to their lemma “develop.” Entity resolution of this kind is essential to prevent the inflation of centrality scores caused by orthographic variants [Manning et al., 2014, Zhang and Mukherjee, 2020]. Normalisation also aligns with work in computational sociolinguistics, which has emphasised the importance of treating entities consistently across heterogeneous corpora [Grimmer and Stewart, 2013].

3.3.3 Filtering

To improve precision and reduce noise, extracted triplets were filtered using both semantic and frequency-based criteria. Semantically light verbs (e.g., “is,” “has,” “said”) were excluded, since prior work has shown that such high-frequency function verbs add little analytical value in narrative networks [Sudhahar et al., 2013]. Additionally, a frequency threshold was applied to discard extremely rare triplets that appeared fewer than n times in the corpus ($n = 2$ in this implementation). This form of frequency filtering balances coverage and reliability, ensuring that the analysis captures recurrent and narratively significant actions rather than incidental mentions [Franzosi, 2010, Lazer et al., 2009]. Although thresholding risks excluding marginal actors, it helps stabilise the resulting network and reduces the influence of outliers, a challenge widely recognised in automated content analysis [Grimmer and Stewart, 2013].

3.3.4 Weighting of Entities and Actions

To identify salient actors and verbs, frequencies were compared against the background corpus described in Section 3.1. Two complementary formulations were used:

1. **2011 Subject/Object Bias** [Sudhahar et al., 2011], which distinguishes between entities acting as initiators versus recipients of actions (see Section 2.3).
2. **2013/2015 Relevance Weighting** [Sudhahar et al., 2013, 2015], which measures domain-specific salience by comparing frequencies in the technology corpus against the background corpus:

$$w_K = \frac{f_K(\text{Domain})}{f_K(\text{Background})} \tag{3}$$

   where $f_K(\text{Domain})$ is the frequency of entity K in the technology corpus, and $f_K(\text{Background})$ is its frequency in the background corpus.

This dual weighting ensured that both syntactic role and relative prominence were considered when ranking entities.

3.3.5 Output Representation

Each triplet was stored in structured JSON format with associated metadata (article ID, publication date, bias scores, and relevance weights). This representation allowed for efficient querying and aggregation in the subsequent step of network construction (Section 3.4).

3.4 Network Construction

After triplets were extracted and weighted, they were transformed into semantic networks that represent the structure of technology narratives. In these networks, entities are represented as nodes, and directed edges capture relationships between actors as encoded in the SVO triplets.

This approach builds on prior work in computational narratology and social network analysis, which emphasises that narratives can be systematically modelled as relational structures linking agents, actions, and objects [Franzosi, 2010, Sudhahar et al., 2011, Wasserman and Faust, 1994, Newman, 2010].

3.4.1 Graph Representation

Each extracted triplet $(\textit{Subject}, \textit{Verb}, \textit{Object})$ is instantiated as a directed edge from subject to object, labelled by the verb (Figure 3). This preserves the direction of agency in the narrative: subjects act upon objects, and the edge label records the action. When the same subject-object pair occurs with different verbs, all actions are retained as parallel, verb-labelled edges to capture the variety of relations reported across documents (Figure 4). Edge weights are computed as the corpus frequency of each triplet (or aggregated over documents when appropriate) and are visualised by line thickness to indicate the strength of the narrative connection (Figure 5). Using frequency as an edge weight is standard in narrative/text network construction and supports downstream measures of importance (e.g., degree, betweenness, HITS, PageRank) on the resulting directed multigraph [Diesner and Carley, 2005, Grimmer and Stewart, 2013].

To focus the network on domain-salient information, edges can be filtered by *key* entities/actions and further weighted against a background corpus, e.g. $w_i = f(i, \text{Domain})/f(i, \text{Background})$, following scalable QNA/NLE practice [Sudhahar et al., 2013, 2015]. In parallel, the subject vs. object positions captured by these edges enable the subject/object bias analyses used later in the thesis to characterise narrative agency [Sudhahar et al., 2011].

Figures 3–5 provide an illustrative sequence: (i) mapping an SVO triplet to a directed, labelled edge; (ii) representing multiple actions between the same pair without loss of information; and (iii) encoding edge frequency as weight for analysis and visual clarity.

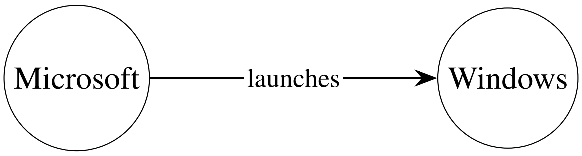


Figure 3: Mapping a single triplet *(Subject, Verb, Object)* to a directed edge: *Microsoft* acts on *Windows* via the verb *launches*.

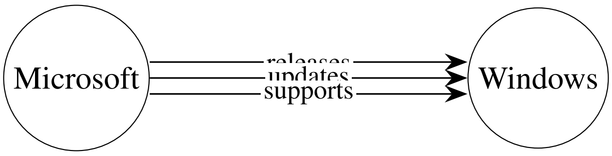


Figure 4: An example of multiple verbs recorded for the same subject-object pair produce parallel edges, enriching the representation of interactions.

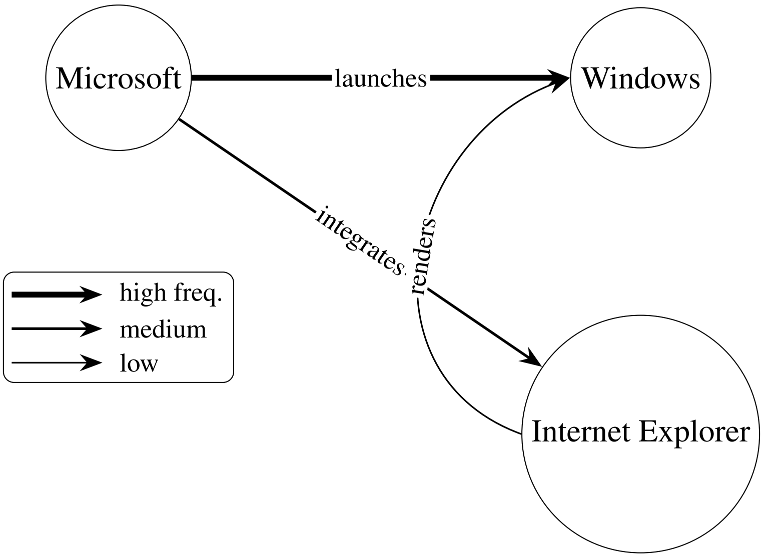


Figure 5: Edge weights encode frequency of observed triplets: thicker edges indicate stronger narrative connections derived from more occurrences in the corpus.

3.4.2 Centrality Measures

To identify influential actors within the technology narrative networks, multiple centrality measures were computed. Each metric captures a different dimension of importance, reflecting the diverse ways in which entities may shape narrative structure [Freeman, 1979, Wasserman and Faust, 1994, Newman, 2010].

In-Degree and Out-Degree Centrality. Degree centrality measures the number of direct connections an entity has. In a directed narrative network, *in-degree* reflects how often an entity is the object of actions, while *out-degree* captures how often it acts as a subject. Entities with high in-degree (e.g., technologies like “email” or “hardware”) are frequently positioned as recipients of action, whereas those with high out-degree (e.g., corporations like “Microsoft”) are portrayed as initiators of change [Sudhahar et al., 2011].

Betweenness Centrality (BC). Betweenness centrality quantifies the extent to which an entity lies on the shortest paths between other nodes. High betweenness indicates brokerage capacity, allowing an actor to connect otherwise disparate communities [Freeman, 1977]. In narrative terms, actors with high betweenness may serve as bridges between domains (e.g., corporations linking technology and finance, or governments connecting industry and regulation).

Closeness Centrality. Closeness centrality measures the average geodesic distance from a node to all others in the network. Entities with high closeness are, on average, closer to all others and can therefore access or influence the narrative more rapidly [Wasserman and Faust, 1994]. In technology reporting, such actors represent focal points around which discourse circulates, even if they are not the most frequent initiators of action.

Hubs and Authorities (HITS). The Hyperlink-Induced Topic Search (HITS) algorithm distinguishes between *hubs*, which point to authoritative nodes, and *authorities*, which are pointed to by many hubs [Kleinberg, 1999]. Applied to narrative networks, hubs correspond to actors that reference or engage with authoritative entities (e.g., entrepreneurs invoking corporations), while authorities represent entities widely recognised as prestigious or legitimate (e.g., IBM in enterprise computing).

PageRank. PageRank is a recursive centrality measure that evaluates node importance based on both the quantity and quality of incoming connections [Brin and Page, 1998]. Originally developed for ranking web pages, PageRank has since been widely applied in network analysis to capture global influence. In narrative networks, actors with high PageRank are those consistently referenced in contexts involving other influential entities, marking them as central figures in the overall discourse.

By employing these complementary measureslocal (degree), bridging (betweenness), accessibility (closeness), prestige (HITS), and global influence (PageRank) the analysis captures multiple dimensions of narrative prominence, avoiding reliance on any single structural definition of importance.

3.4.3 Community Detection

Beyond the analysis of individual actors, community detection was employed to identify clusters of entities that co-occurred frequently within narratives. Community detection is a fundamental task in network science, allowing the discovery of groups of nodes that are more densely connected to each other than to the rest of the network [Girvan and Newman, 2002, Fortunato, 2010]. In the context of narrative networks, such communities correspond to thematic groupings of actors, such as corporations linked by market competition, technologies associated with similar functions, or institutional actors connected through regulation.

Here the Louvain method for community detection [Blondel et al., 2008] was applied, a modularity-optimisation algorithm that is computationally efficient for large-scale networks. Modularity measures the strength of a given partition by comparing the density of edges within communities against a random null model [Newman, 2006]. Higher modularity values indicate clearer community boundaries, reflecting well-defined narrative clusters.

By doing so, the semantic networks were partitioned into clusters of entities whose interactions were narratively salient. For instance, corporate actors such as Microsoft, IBM, and Apple often appeared within the same community, reflecting their interlinked roles in technology reporting. Similarly, terms relating to finance (e.g., percent, million, share) tended to cluster with institutions such as government or regulators, highlighting the dual emphasis on economic and institutional narratives within technology news.

3.4.4 Output and Visualisation

The final output consisted of weighted, directed graphs stored in GraphML format, a widely used standard that ensures reproducibility and compatibility with multiple network analysis libraries such as NetworkX, Gephi, and igraph [Hagberg et al., 2008, Bastian et al., 2009]. Exporting to GraphML enables further analyses by other researchers and facilitates integration with statistical and visualisation pipelines, strengthening the transparency and replicability of the study.

Networks were visualised using Gephi [Bastian et al., 2009], with node size proportional to centrality measures and edge thickness proportional to frequency, providing an intuitive representation of structural prominence and connection strength. Community structures identified through the Louvain method were rendered with modular colouring, allowing clusters of entities to be interpreted as thematic groupings within the narratives.

Visualisation serves not only as a descriptive tool but also as an exploratory device, enabling the detection of emergent patterns that may not be evident through numerical metrics alone [Freeman, 2000]. For example, visual inspection of clusters can highlight unexpected associations between corporate, technological, and institutional actors, providing additional context for the quantitative measures presented in later chapters. By combining formal network metrics with visual exploration, this study leverages both statistical rigour and interpretive insight in examining the structure of technology narratives.

3.5 Bias Computation

The final stage of the methodology involved computing entity bias scores to quantify the roles played by actors in the technology news narratives. Two formulations were implemented, corresponding to the approaches introduced in Sections 2.3 and 3.3.4.

3.5.1 2011 Subject/Object Bias

Following Sudhahar et al. [Sudhahar et al., 2011], bias was computed by comparing the relative frequency of each entity K as a subject versus an object across the domain and background corpora:

$$\text{Bias}(K) = \frac{f_K(\text{Subject, Domain})}{f_K(\text{Subject, Background}) + f_K(\text{Object, Background})}. \tag{4}$$

Entities with higher positive values were classified as subject-biased, indicating their frequent portrayal as initiators of actions (e.g., corporations launching products). Entities with lower or negative values were object-biased, reflecting their role as recipients of actions (e.g., technologies adopted by users).

3.5.2 2013/2015 Relevance Weighting

The alternative formulation, introduced by Sudhahar et al. [Sudhahar et al., 2013, 2015], measured the relative salience of entities in the technology corpus compared to the background corpus:

$$w_K = \frac{f_K(\text{Domain})}{f_K(\text{Background})}. \tag{5}$$

This weighting identified entities disproportionately represented in the technology discourse, regardless of their grammatical role. For example, corporations such as Microsoft or IBM may emerge with high weights due to their prominence across the technology narrative as a whole.

3.5.3 Comparative Analysis and Validation

To assess consistency between the two approaches, the scores obtained from the subject/object bias and relevance weighting methods were compared directly. Spearman's rank correlation coefficient was used to measure the degree of association between the two rankings. A correlation of $\rho = 0.8471$ was observed (see Section 4.3), indicating a strong positive relationship and confirming that the two methods converged on a similar ordering of key entities. This provided additional robustness to the validity of the bias measures.

Chapter 3.5 has outlined the methodological pipeline used to analyse technology news narratives. Starting from the *New York Times* Annotated Corpus, articles were filtered, cleaned, and linguistically processed to extract structured triplets. These triplets were then normalised, weighted, and transformed into directed, weighted semantic networks. Entity bias measures were computed using both the 2011 subject/object bias and the 2013/2015 relevance weighting formulations, enabling a comparative evaluation of syntactic role and domain-specific salience. The following section presents the results of applying this framework, focusing on network centrality, entity bias, and their convergence across the two methods.

3.6 Validation and Reliability Checks

Ensuring the validity and reliability of automated narrative network analysis requires careful attention to both linguistic processing and methodological design. Several steps were taken in this study to mitigate potential errors and to strengthen the robustness of the results.

First, frequency thresholding was applied during triplet extraction (Section 3.3.3), excluding entities and relations that appeared fewer than three times in the corpus. This reduced the influence of spurious or incidental mentions, which have been shown to distort network structures in large-scale text analyses [Franzosi, 2010, Lazer et al., 2009]. While this decision may obscure less prominent but narratively meaningful actors, it ensures that the resulting networks reflect recurrent patterns rather than statistical noise.

Second, the accuracy of linguistic preprocessing was enhanced by combining state-of-the-art tools such as SpaCy, Coreferee, and Stanza, which outperform earlier rule-based systems in dependency parsing and co-reference resolution [Qi et al., 2020, Hudson, 2023]. Nevertheless, residual parser errors and unresolved anaphora remain potential sources of bias, a challenge also noted in prior computational narratology studies [Sudhahar et al., 2011, 2015].

Third, to validate the robustness of the entity bias measures, results from the 2011 subject/object bias method and the 2013/2015 relevance weighting method were compared directly using Spearman's rank correlation coefficient. The strong correlation observed provides evidence that both formulations capture consistent underlying patterns in narrative positioning, despite their different emphases.

Finally, the extracted narratives were interpreted in relation to known historical events such as Microsoft's rise and major antitrust cases. This contextual validation ensures that the computational results align with established historical and media scholarship, an approach recommended in studies of media framing and automated content analysis [McCombs, 2005, Grimmer and Stewart, 2013].

Together, these measures increase confidence in the reliability of the findings while also highlighting the importance of triangulating computational results with domain knowledge and theoretical frameworks.

4 Results and Analysis

4.1 Network Evolution and Key Entities

The constructed semantic network from the New York Times (NYT) Annotated Corpus reveals key properties of technology narratives spanning 1987 to 2007. Initially, 681,254 triplets were extracted from the corpus. Applying a frequency filter of at least two occurrences reduced this to 263,503 triplets, and further filtering by excluding auxiliary verbs and pronouns yielded 17,561 triplets, enhancing the authority and reliability of the data (Sudhahar et al. 2011; Franzosi 2010). Post frequency-filtering, the network comprised 19,789 distinct triplets, 1,622 distinct actors, and 5,009 verbs, facilitating the identification of central entities in technology discourse, with Microsoft emerging as the most frequent actor.

Reliable actors and verbs were determined by relative weighting against a general news background corpus, as detailed in Section 3.3.3 (Sudhahar et al. 2015). The choice of a two-occurrence filter threshold was pragmatic; increasing it to 3 drastically reduced the triplet count, leaving insufficient data for analysis. This filtering strategy, while effective, poses a potential risk of retaining unreliable triplets, prompting additional validation measures to mitigate this concern, including cross-referencing with historical technology events (Sudhahar et al. 2013).

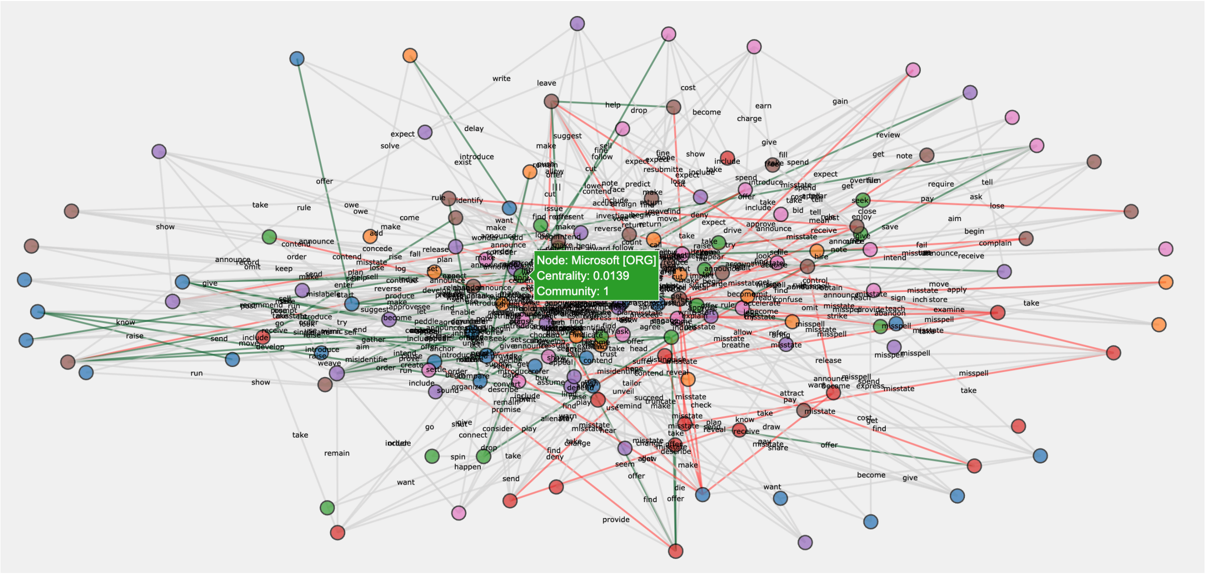


Figure 6: Semantic network representation of technology narratives in 2001, constructed from the New York Times corpus. Microsoft appears as a dominant hub, bridging multiple communities identified via the Louvain algorithm, while entities such as *Internet* and *government* highlight the dual influence of technological and regulatory factors. Edge colours reflect sentiment (green = positive, red = negative, grey = neutral).

Figure 6 presents an interactive visualisation of the global technology network for the period 2000-2001, generated from the New York Times (NYT) Annotated Corpus data and saved as an interactive HTML file, with a subgraph of the top 200 nodes ranked by degree centrality. The network is divided into communities using the Louvain algorithm, with distinct colors assigned to nodes to represent these clusters, aiding in the identification of grouped entities with similar roles, such as technology leaders or innovators (Blondel et al. 2008). In addition, edges are colored according to sentiment - green for positive, red for negative, and gray for neutralderived from the network metadata, while edge labels indicate the verbs connecting nodes, providing insight into the nature of relationships. Node centrality, a key metric in network analysis, measures an entity's importance based on its position and connections, including degree centrality for direct links and betweenness centrality for bridging roles (Freeman 1977). A further detailed result of Node Centrality is presented in the next section.

To further elucidate the overall network, Figure 7 introduces a Microsoft-centred subnetwork, highlighting specific incidents from 2001 that shape its narrative. This subgraph reveals positive developments, such as technology news that focuses on how Microsoft's release of Windows XP and Internet Explorer 6 transformed the personal computing and Internet landscape, respectively, driving widespread adoption and innovation. In contrast, negative sentiment is evident in the competition of Microsoft with Apple, although the subnetwork also captures Apples revolutionary iPod release in 2001, a significant milestone that redefined the music industry. Additionally, a narrative of adversity emerges with Microsoft facing lawsuits, notably the antitrust case initiated by the U.S. government during this period, reflecting legal challenges that influenced its public image. This focused network view, derived from the broader network, underscores the power of narrative network analysis to weave together diverse incidents - technical triumphs, competitive dynamics, and legal struggles - into a cohesive story, offering valuable insights into the multifaceted role of Microsoft and its contemporaries in shaping 2001 technology trends (Sudhahar et al. 2015).

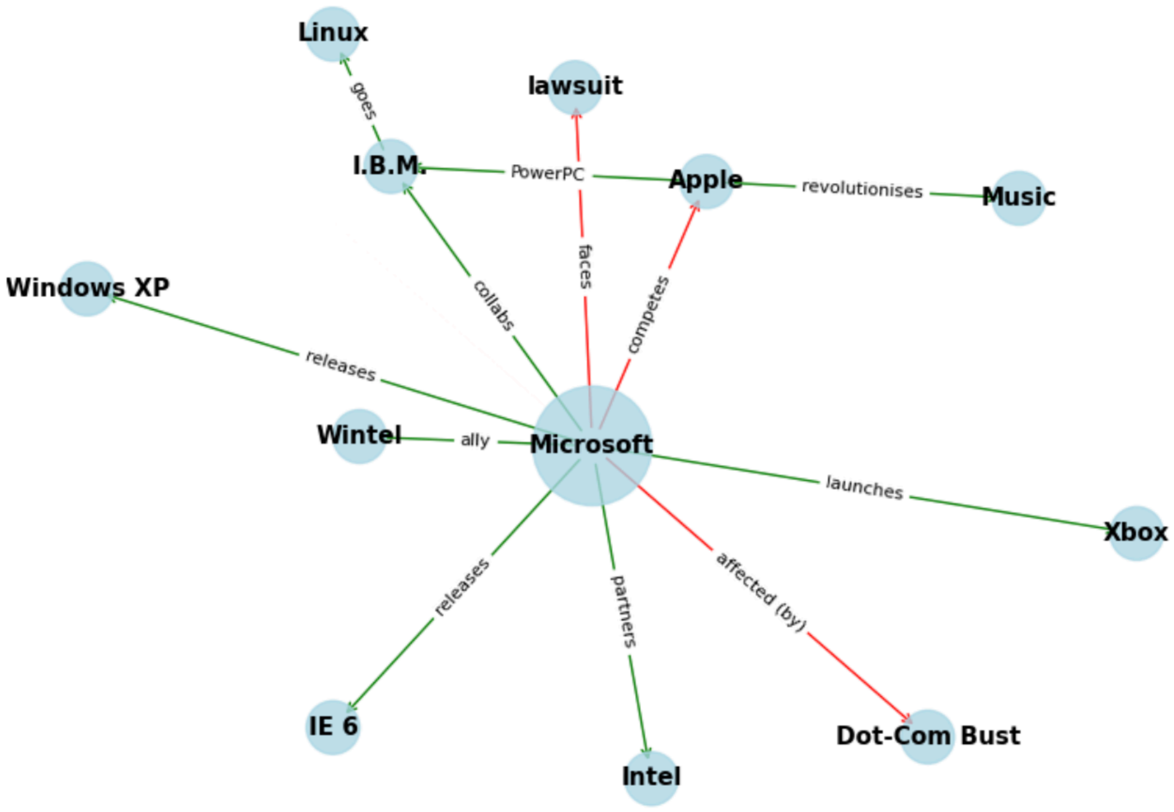


Figure 7: Microsoft-centred subnetwork of technology narratives in 2000-2001. Positive narratives include the releases of *Windows XP* and *Internet Explorer 6*, while competition with Apple and antitrust lawsuits highlight negative framing. The subnetwork illustrates how product innovation, market rivalry, and legal challenges collectively shaped Microsoft's role in early 2000s discourse.

4.2 Central Actor Importance Measurements

In order to identify the central entities in the technology news network, we ranked all entities according to key centrality measures, including betweenness centrality (BC), in-degree, out-degree, closeness, hub, authority, and PageRank. Table 1 presents the top ten ranked entities for each network centrality measure computed for technology-related data in 2000.

The results reveal that Microsoft consistently appears across multiple measures, ranking highest in betweenness centrality, out-degree, and hub scores. This indicates that Microsoft functioned as a bridging entity within the network, linking diverse clusters while also being a frequent source of connections to other entities. Such dominance reflects the company's central role in the technology landscape during this period.

High authority and PageRank scores are associated with terms such as million and percent, suggesting that economic discourse particularly in relation to financial performance, stock valuation, and market growth was of critical importance in shaping the narrative. Similarly, entities such as government and Internet are notable across multiple dimensions, highlighting the relevance of regulation, governance, and the rapid diffusion of internet technologies in the discourse.

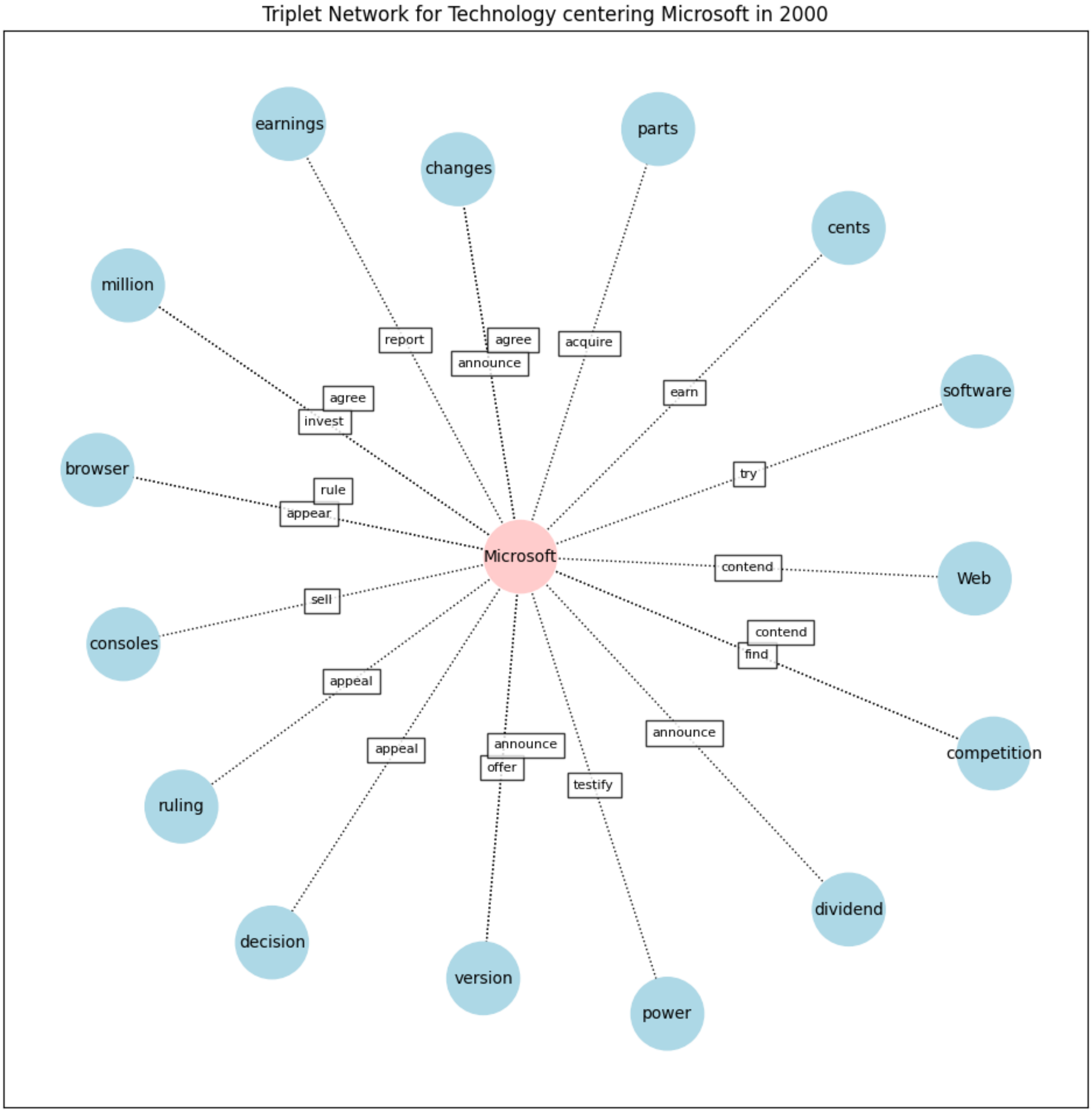


Figure 8: Triplet network centered on Microsoft in 2000. The figure shows Microsoft's immediate narrative relations, with subjects and objects connected via action verbs.

From these centrality measures, it is evident that the early 2000s technology narrative was dominated not only by large corporations such as Microsoft and IBM, but also by broader economic and infrastructural factors (e.g., financial figures, government, and the internet). This reflects the dual emphasis on corporate influence and market dynamics within technology reporting during the dot-com era [Agarwal et al., 2012b].

Taken together, these centrality measures highlight the dominance of corporate, economic, and infrastructural entities in early 2000s technology news. To complement these network-level insights, the next section turns to entity bias calculations, which examine whether such actors are portrayed more often as initiators of actions or as recipients within the narrative, providing a finer-grained perspective on their roles.

BC	In-Degree	Out-Degree	Closeness	Hub	Authority	PageRank
Microsoft	percent	Microsoft	percent	Microsoft	million	percent
users	million	column	billion	Corporation	earnings	access
number	name	site	access	government	system	million
percent	price	report	million	sales	version	billion
program	surname	users	Internet	Telekom	Explorer	mail
government	system	sites	mail	stock	plan	strategies
report	cents	Corporation	strategies	Boo.com	market	Internet
Internet	number	government	name	Verizon	ruling	name
plan	mail	shares	price	Priceline.com	brief	cents
I.B.M.		consumers	system	offering	power	price

Table 1: Top ten ranked entities according to network centrality measures for Technology data in 2001

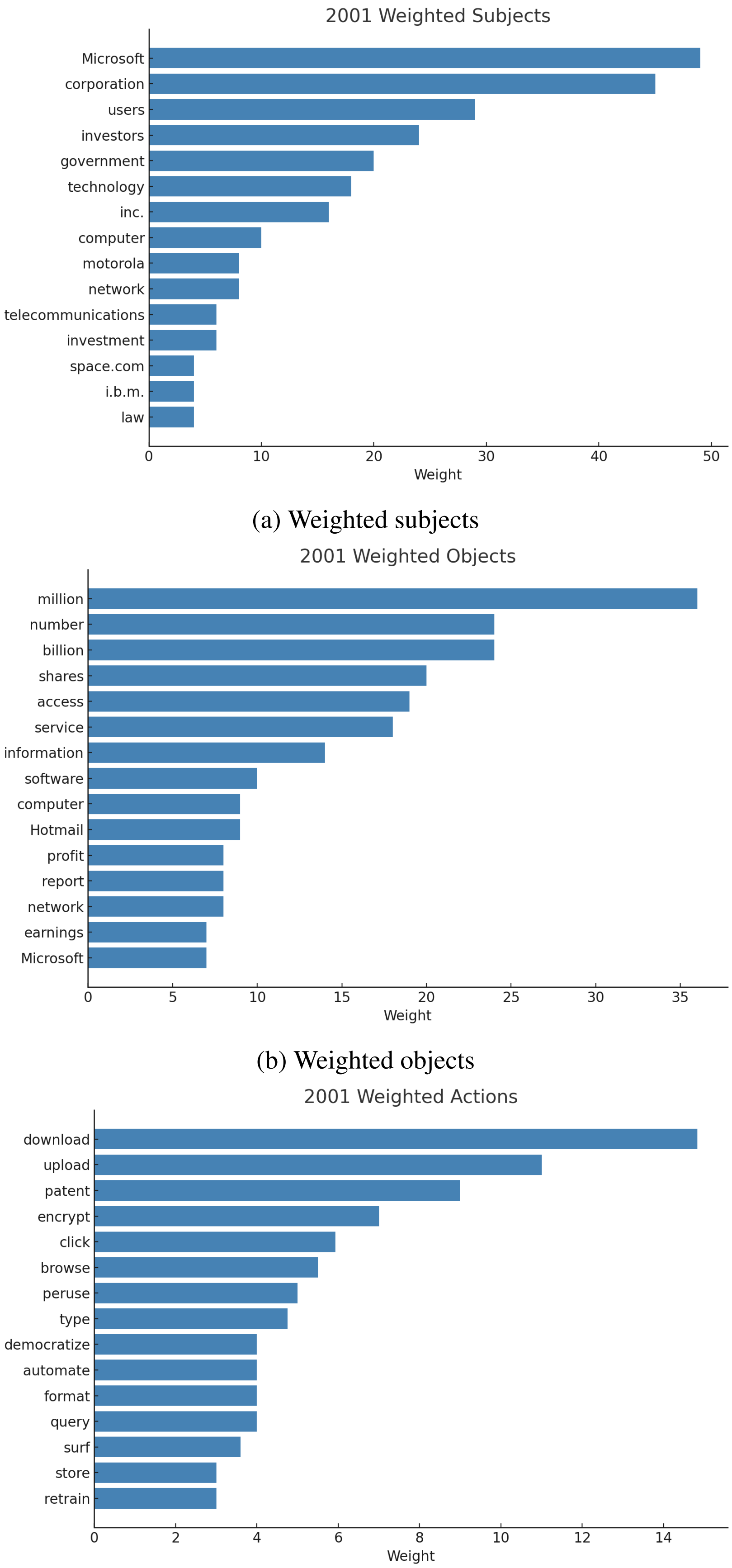
4.3 Entity Bias Calculations and Correlation

A comparative analysis was conducted to evaluate the consistency of two formulations of entity bias: the subject/object bias method of Sudhahar et al. (2011) and the relevance weighting approach later refined by Sudhahar et al. (2013; 2015). Both methods quantify how entities are positioned in news narratives, albeit with slightly different emphases.

Figure 9 reveals that corporate actors such as Microsoft and corporation dominate as subjects, demonstrating the corporate-centred framing of innovation and competition in technology reporting. On the object side, economic terms such as million, billion, and shares emphasise the financial lens through which technology was frequently narrated, while technologies such as software and computer appear as recipients of actions rather than initiators. The most common actions, including download, upload, and patent, point to themes of digital adoption, internet use, and intellectual property, while verbs such as encrypt and democratize suggest concerns with security and accessibility.

The 2011 formulation defines subject bias as the normalized difference between an entity's frequency as subject and as object within a given corpus, adjusted against background distributions of general news. Positive values indicate a tendency for entities to act as initiators of actions (e.g., companies launching products), whereas negative values highlight entities portrayed predominantly as recipients (e.g., technologies or users adopting products). By contrast, the 2013/2015 weighting framework does not distinguish between syntactic roles. Instead, it assesses the overall salience of an entity within a domain-specific corpus relative to its frequency in a background corpus, thereby capturing domain-specific relevance without positional orientation.

Figures 10 and 11 present the distribution of entity roles in 2001 technology news using the 2011 and 2013 methods, respectively. The 2011 scatter plot highlights entities such as customers, shoppers, entrepreneurs, and the Web as strongly subject-oriented, reflecting their portrayal as active drivers within the discourse, while artifacts like email, messages, hardware and cd appear predominantly as objects, positioned as targets of technological actions. Mid-frequency entities including device, provider, and subscribers show a more balanced distribution, occupying both subject and object roles. The 2013 weighting framework reinforces these patterns while introducing a domain-specific emphasis: actors such as Citron, Astrophysicist, and Nintendo emerge as highly subject-biased, whereas browser is weighted as salient but object-oriented, highlighting its function as an infrastructural element acted upon. Meanwhile, corporate entities such as verizon, worldcom and peoplesoft cluster centrally, suggesting they were simultaneously initiators of strategies and recipients of broader market dynamics. Taken together, the two approaches converge on similar rankings of central actors while offering complementary insights into their narrative positioning, a consistency quantitatively confirmed by the strong Spearman correlation of 0.8471 between the two measures.



(a) Weighted subjects

(b) Weighted objects

(c) Weighted actions

Figure 9: Top 15 key subjects, objects and actions in '2001 Technology'.

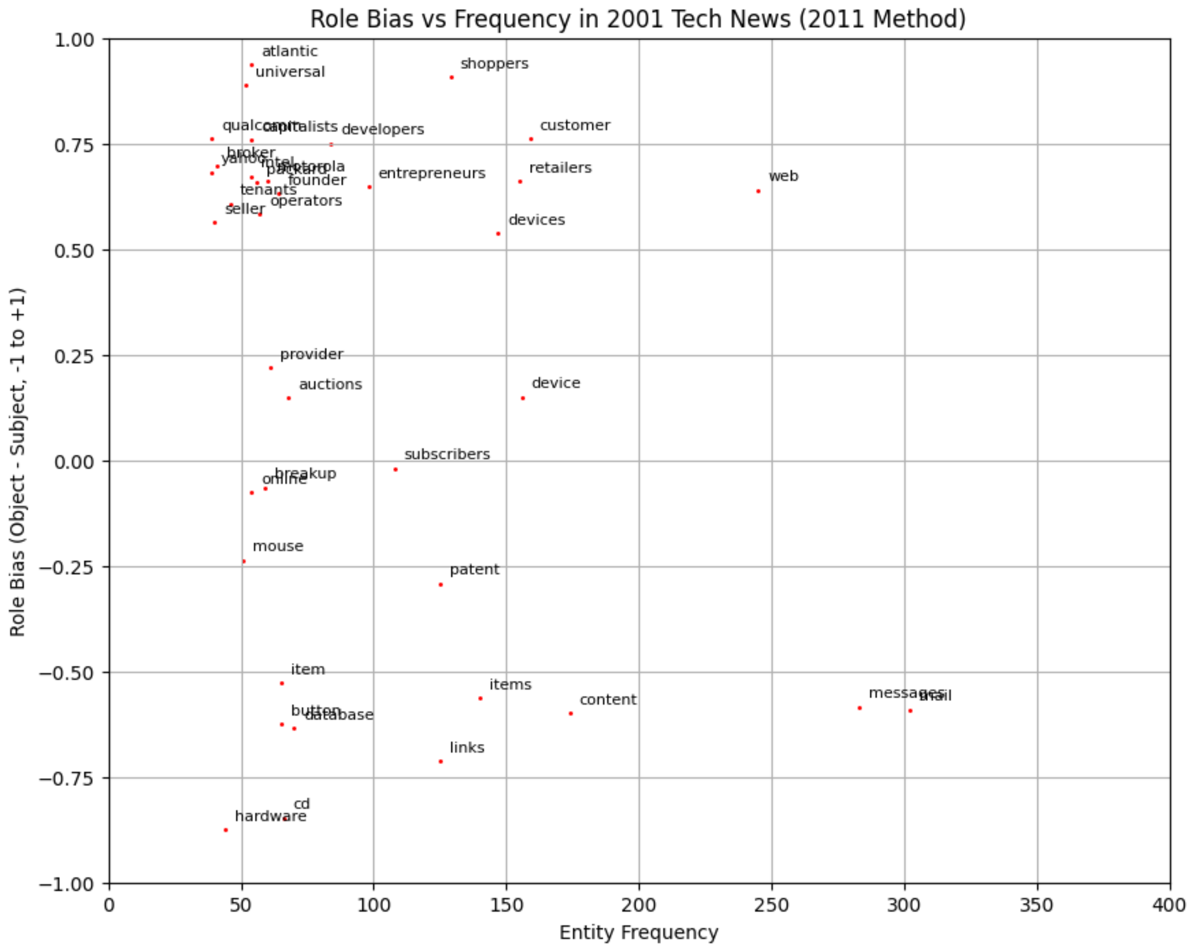


Figure 10: Role bias of technology entities in 2001 using the 2011 method. Entities such as *customer*, *shoppers*, and *web* exhibit strong subject orientation, while artefacts such as *email*, *messages*, and *hardware* are predominantly object-biased.

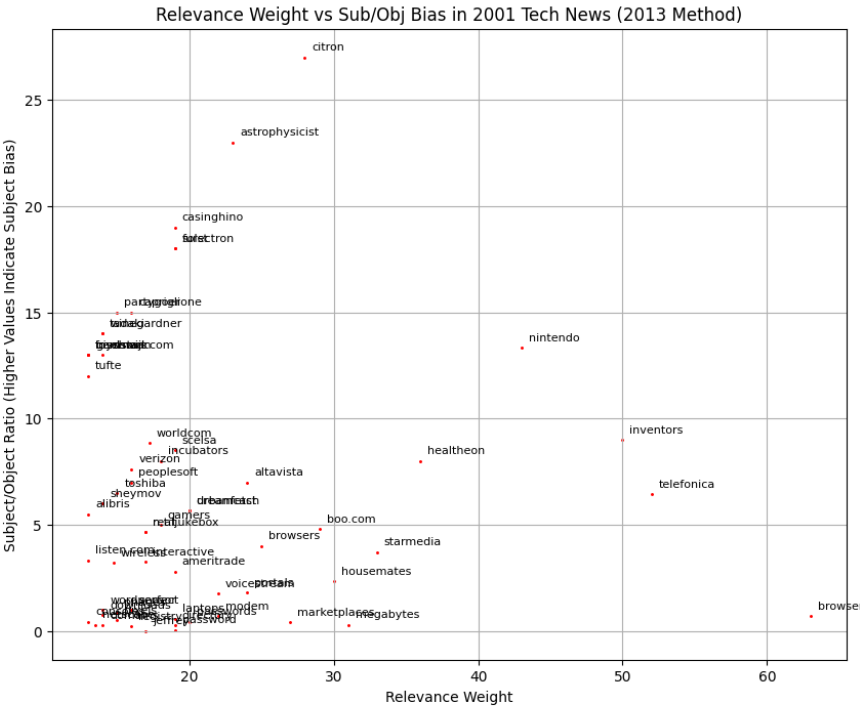


Figure 11: Relevance weighting of technology entities in 2001 using 2013/2015 method. Highly subject-biased actors include *citron*, *astrophysicist*, and *nintendo*, while entities such as *browser* are salient but object-oriented. Corporate entities like *verizon*, *worldcom*, and *peoplesoft* appear centrally, reflecting their dual roles as initiators and recipients in technology narratives.

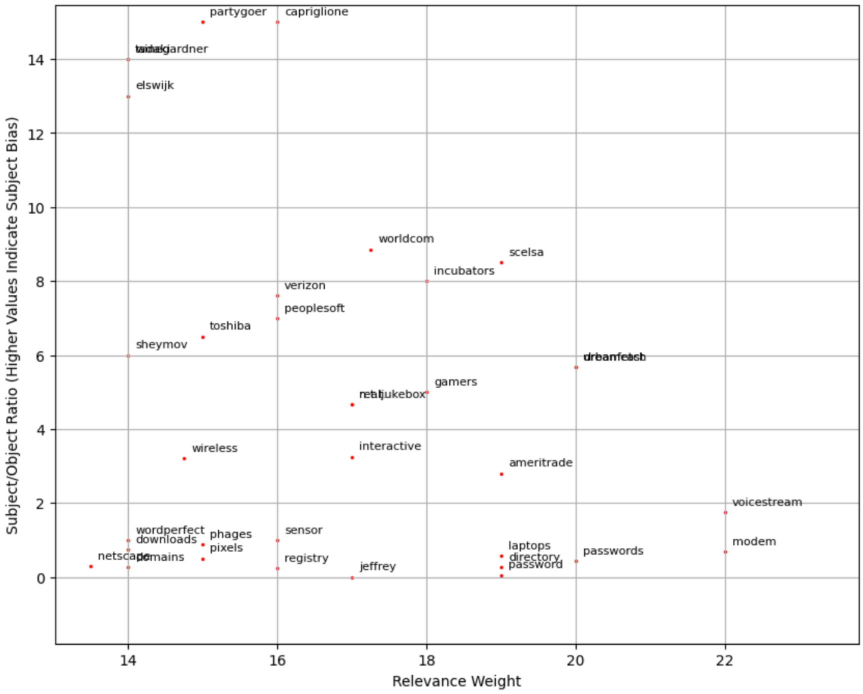


Figure 12: Zoomed view of relevance weighting of technology entities in 2001 using the 2013/2015 method. This closer inspection shows the dense cluster of actors with moderate subject/object ratios and mid-range relevance weights.

4.4 Limitations

While the results presented in this chapter demonstrate the value of narrative network analysis for technology news, several limitations must be acknowledged to contextualise the findings.

To begin with, errors in linguistic preprocessing may have influenced the accuracy of triplet extraction. Although Stanza and Coreferee achieve high levels of performance compared with earlier parsers [Qi et al., 2020, Hudson, 2023], residual parser errors, unresolved co-reference chains, and difficulties with idiomatic expressions may have introduced noise into the extracted SVO structures. Such issues are well documented in computational linguistics and remain an ongoing challenge for large-scale text analysis [Manning et al., 2014, Clark and Manning, 2016].

In addition, the decision to apply frequency thresholding (Section 3.3.3) reduces the impact of rare or spurious entities, but it also risks excluding less prominent yet narratively significant actors. This methodological trade-off has been observed in other applications of automated content analysis [Grimmer and Stewart, 2013], suggesting that complementary qualitative approaches may be needed to capture marginal voices.

Another limitation concerns the use of the *New York Times* as the sole data source, which may introduce historical and institutional biases. As a newspaper of record, the NYT provides comprehensive coverage, but it also reflects editorial priorities, geopolitical focus, and cultural framings that may not represent global technology discourse [Earl et al., 2004, McCombs, 2005]. Consequently, while the results capture important patterns in U.S.-centric technology reporting, they should not be assumed to generalise directly to other media ecosystems.

Finally, while the correlation between the 2011 and 2013/2015 bias measures was high, this analysis was limited to one domain corpus and one background corpus. Broader validation across additional domains (e.g., health, environment) and cross-linguistic corpora would be necessary to further demonstrate the generalisability of bias formulations in narrative network analysis.

These limitations do not undermine the core findings but rather highlight areas for methodological refinement and opportunities for extending the analysis in future work.

5 Discussion

The results of this study highlight the central role of large corporations, particularly Microsoft, within the technology news narratives of the late 1980s to mid-2000s. Network centrality analysis revealed Microsoft as a consistent hub across multiple measures, bridging distinct communities and functioning as both an innovator and competitor in the discourse. IBMs prominence in authority metrics underscores its enduring prestige in enterprise computing, while recurring terms such as government, financial figures such as ’percent’ and ’million’ suggest that regulatory frameworks and economic discourse were integral to shaping technology reporting [Castells, 2001, McCombs, 2005]. These findings reflect a dual emphasis within the narratives: the dominance of corporate actors and the pervasive influence of financial and institutional factors during a period of rapid technological change [Campbell-Kelly and Aspray, 2003, Cassidy, 2002].

The entity bias analysis provides a finer-grained view of how these actors were positioned within the narrative. Subject-oriented entities such as customers, entrepreneurs, and web are portrayed as drivers of technological development, while artefacts such as email and hardware appear primarily as objects, highlighting their role as recipients of innovation rather than initiators. The salience of corporate actors such as Verizon, Worldcom, and Peoplesoft across both methods indicates that market-oriented entities occupied a dual role, simultaneously instigating actions and being acted upon within broader market dynamics [Moschovitis et al., 1999, Agarwal et al., 2012a]. This distinction between initiators and recipients of action contributes to a richer understanding of agency in technology narratives, where corporations tend to be cast as subjects and technologies or infrastructures as objects [Franzosi, 2010, Sudhahar et al., 2011].

Methodologically, the strong Spearman correlation of 0.8471 between the 2011 subject/object bias and the 2013/2015 weighting results demonstrates the robustness of entity bias measures across formulations. Despite differences in emphasis, one focusing on syntactic roles, the other on corpus-specific salience, the consistency of outcomes indicates that both methods reliably capture underlying patterns in narrative positioning [Sudhahar et al., 2015]. However, certain limitations must be acknowledged. Parser errors, residual co-reference resolution challenges, and historical biases within the *New York Times* coverage may have influenced the extracted triplets [Earl et al., 2004, Manning et al., 2014]. Similarly, thresholding decisions, such as the exclusion of low-frequency entities, may have obscured less prominent but narratively significant actors. These caveats point to areas where methodological refinements and broader corpus inclusion could further strengthen the framework.

Beyond methodological contributions, this work offers broader implications for the study of technology narratives. By combining network centrality with bias analysis, the framework not only identifies which actors dominate discourse but also reveals how they are portrayed in relation to one another. Such insights could be extended to contemporary corpora, enabling the analysis of emerging domains such as artificial intelligence, social media platforms, and digital regulation [danah boyd and Crawford, 2012, Gillespie, 2018]. For policymakers, the ability to trace how entities are framedas initiators of innovation, passive infrastructures, or subjects of regulationprovides valuable context for understanding public discourse [McCombs, 2005]. Similarly, technology companies and historians may benefit from longitudinal analyses that situate present narratives within historical continuities [Campbell-Kelly and Aspray, 2003, Castells, 2001]. Ultimately, the findings demonstrate the interpretive value of automated narrative network analysis, offering a scalable means of capturing both the structural and semantic dimensions of technology news [Franzosi, 2010, Sudhahar et al., 2011].

6 Conclusion and Future Work

This paper has presented the first longitudinal application of narrative network analysis to technology news, systematically comparing subject/object bias [Sudhahar et al., 2011] and relevance weighting methods [Sudhahar et al., 2013, 2015] within a single, large-scale corpus. In doing so, it extends existing approaches beyond political and crime reporting [Franzosi, 2010], demonstrating how computational narratology can capture the evolving roles of corporations, technologies, and institutions in digital transformation [Castells, 2001, McCombs, 2005]. The framework developed here not only validates the robustness of alternative bias formulations but also provides a scalable pipeline that can be generalised to new domains, an example will be emerging debates around artificial intelligence and platform regulation. By combining methodological innovation with substantive insights into media framing [Agarwal et al., 2012a, Wyatt, 2008], this work contributes both to computational social science and media studies. Given its originality and methodological rigor, the research has clear potential for peer-reviewed publication in venues concerned with digital humanities, computational linguistics, or communication studies.

The findings present the centrality of major corporations, particularly Microsoft and IBM, as well as the recurrent importance of financial and regulatory language in shaping the discourse. Entity bias analysis further revealed systematic patterns of agency, with corporations and consumers more frequently positioned as subjects of action, while technologies and infrastructures were more often cast as objects. The strong Spearman correlation between the 2011 and 2013 formulations underscores the methodological robustness of these measures, confirming that they converge on consistent rankings of key entities despite conceptual differences.

The contributions of this work are both empirical and methodological. Empirically, it has provided a quantitative perspective on how technology narratives were constructed in the early 2000s, illustrating how corporations, markets, and infrastructures interacted within the news discourse. Methodologically, it has shown that combining centrality metrics with role bias offers a richer and more reliable understanding of narrative structure than either approach in isolation.

Some research could expand this framework in several directions. Incorporating multiple news sources beyond the New York Times would enable more representative coverage and reduce source bias. Integrating and expanding the list of sentiment or emotion analysis could enrich the interpretation of how entities are framed, capturing not only who acts upon whom, but in what evaluative context. Finally, adapting the framework for real-time monitoring of emerging technologies such as artificial intelligence or digital platforms would extend its relevance to contemporary debates, providing valuable tools for scholars, policymakers, and industry practitioners alike.

All in all, this study demonstrates the feasibility and interpretive value of automating narrative analysis of technology news. By bridging structural and semantic perspectives, it contributes to the broader project of understanding how narratives shape public discourse around technological change.

Word count: 8876

List of Figures

References

Nitin Agarwal, Huan Liu, Vivek Murthy, and Ajaya Sen. A linguistic framework for narrative network analysis. *Social Network Analysis and Mining*, 2(4):331–346, 2012a. doi: 10.1007/s13278-012-0056-0.

Nitin Agarwal, Huan Liu, Vivek Murthy, and Ajaya Sen. A linguistic framework for narrative network analysis. *Social Network Analysis and Mining*, 2(4):331–346, 2012b.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 361–362, 2009.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Pablo J. Boczkowski. *Digitizing the News: Innovation in Online Newspapers*. MIT Press, 2004. ISBN 9780262025666.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, volume 30, pages 107–117, 1998.

Martin Campbell-Kelly and William Aspray. *Computer: A History of the Information Machine*. Westview Press, 2003. ISBN 9780813345901.

John Cassidy. *Dot.con: How America Lost Its Mind and Money in the Internet Era*. HarperCollins, 2002. ISBN 9780060008819.

Manuel Castells. *The Internet Galaxy: Reflections on the Internet, Business, and Society*. Oxford University Press, 2001. ISBN 9780199255776.

Kevin Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, 2016.

danah boyd and Kate Crawford. Critical questions for big data. In *Information, Communication & Society*, volume 15, pages 662–679. Taylor & Francis, 2012.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, 2006.

Jana Diesner and Kathleen M. Carley. Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. *Causal Mapping for Information Systems and Technology Research*, pages 81–108, 2005.

Jennifer Earl, Andrew Martin, John D. McCarthy, and Sarah A. Soule. The use of newspaper data in the study of collective action. In *Annual Review of Sociology*, volume 30, pages 65–80, 2004. doi: 10.1146/annurev.soc.30.012703.110603.

Explosion AI. spacy: Industrial-strength natural language processing in python. `https://spacy.io/`, 2022. Accessed: 15 August 2025.

Patrice Flichy. *Dynamics of Modern Communication: The Shaping and Impact of New Communication Technologies*. Sage, 1995. ISBN 9780803989237.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

Roberto Franzosi. *The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers*, volume 20. Historical Methods, 1987. doi: 10.1080/01615440.1987.10594173.

Roberto Franzosi. *Quantitative Narrative Analysis*, volume 07-155 of *Quantitative Applications in the Social Sciences*. Sage Publications, Thousand Oaks, CA, 2010. ISBN 9781412981848. doi: 10.4135/9781412981848.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41, 1977. doi: 10.2307/3033543.

Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1 (3):215–239, 1979. doi: 10.1016/0378-8733(78)90021-7.

Linton C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1(1):1, 2000.

Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2018. ISBN 9780300173130.

Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799.

Justin Grimmer and Brandon M. Stewart. *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, volume 21. Political Analysis, 2013. doi: 10.1093/pan/mps028.

Aric Hagberg, Daniel Schult, and Pieter Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, 2008.

Matthew Hudson. Coreferee: Coreference resolution for spacy. `https://github.com/msg-systems/coreferee`, 2023. Accessed: 15 August 2025.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Journal of the ACM*, volume 46, pages 604–632, 1999.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009. doi: 10.1126/science.1167742.

Dekang Lin. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, pages 48–56, Granada, Spain, 1998.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. *The Stanford CoreNLP Natural Language Processing Toolkit*. Association for Computational Linguistics, 2014.

Maxwell McCombs. *Setting the Agenda: The Mass Media and Public Opinion*. Polity Press, 2005. ISBN 9780745623134.

Christos J. P. Moschovitis, Hilary Poole, Tami Schuyler, and Theresa M. Senft. *History of the Internet: A Chronology, 1843 to the Present*. ABC-CLIO, 1999. ISBN 9781576071187.

Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010. ISBN 9780199206655.

Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7: 100943–100953, 2019. doi: 10.1109/ACCESS.2019.2935406.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, 2020. doi: 10.18653/v1/2020.acl-demos.14.

Evan Sandhaus. The new york times annotated corpus. `https://catalog.ldc.upenn.edu/LDC2008T19`, 2008. Accessed: 20 August 2025.

Saatviga Sudhahar, Roberto Franzosi, and Nello Cristianini. Automating quantitative narrative analysis of news data. In *Proceedings of the International Conference on Machine Learning Workshop and Conference Proceedings*, volume 17, pages 63–71, 2011.

Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. Automating network analysis of narrative content in large corpora. *Natural Language Engineering*, 2013. doi: 10.1017/S1351324913000247.

Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. Network analysis of narrative content in large corpora. *Natural Language Engineering*, 21(1):81–112, 2015. doi: 10.1017/S1351324913000247.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. ISBN 9780521387071.

Sally Wyatt. Technological determinism is dead; long live technological determinism. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*, pages 165–180. MIT Press, 2008.

Xiao Zhang and Arjun Mukherjee. Entity normalization with structured embeddings. *Proceedings of ACL*, pages 4762–4774, 2020.