Big Tech influence over AI research revisited: memetic analysis of attribution of ideas to affiliation

Stanisław Giziński[a], Paulina Kaczyńska[a], Hubert Ruczyński[b], Emilia Wiśnios[a,c], Bartosz Pieliński[d], Przemysław Biecek[a,b], Julian Sienkiewicz[e,*]

[a] *University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Warsaw, Poland*
[b] *Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland*
[c] *NASK National Research Institute, Warsaw, Poland*
[d] *University of Warsaw, Faculty of Political Science and International Studies, Warsaw, Poland*
[e] *Warsaw University of Technology, Faculty of Physics, Warsaw, Poland*

Abstract

There exists a growing discourse around the domination of Big Tech on the landscape of artificial intelligence (AI) research, yet our comprehension of this phenomenon remains cursory. This paper aims to broaden and deepen our understanding of Big Tech's reach and power within AI research. It highlights the dominance not merely in terms of sheer publication volume but rather in the propagation of new ideas or *memes*. Current studies often oversimplify the concept of influence to the share of affiliations in academic papers, typically sourced from limited databases such as arXiv or specific academic conferences.

The main goal of this paper is to unravel the specific nuances of such influence, determining which AI ideas are predominantly driven by Big Tech entities. By employing network and memetic analysis on AI-oriented paper abstracts and their citation network, we are able to grasp a deeper insight into this phenomenon. By utilizing two databases: *OpenAlex* and *S2ORC*, we are able to perform such analysis on a much bigger scale than previous attempts.

Our findings suggest that while Big Tech-affiliated papers are disproportionately more cited in some areas, the most cited papers are those affiliated with both Big Tech and Academia. Focusing on the most contagious memes, their attribution to specific affiliation groups (Big Tech, Academia, mixed af-

*Corresponding author, email: julian.sienkiewicz@pw.edu.pl

filiation) seems equally distributed between those three groups. This suggests that the notion of Big Tech domination over AI research is oversimplified in the discourse.

Keywords: Knowledge Diffusion, Novelty, Affiliation Influence, Big Tech Impact, Complex Networks, Natural Language Processing

1. Introduction

Artificial Intelligence (AI) research is often seen as dominated by Big Tech companies. These companies, with their immense computing resources and access to vast amounts of data, have undoubtedly influenced the development of the field. However, while this heavy influence has been beneficial in certain aspects, it has also raised concerns that the strong position of Big Tech could influence the direction and character of AI, resulting in significant losses for society and science. One of the issues raised is the reproducibility of Big Tech's research (Ebell et al., 2021). Given their proprietary datasets and custom software, it is often difficult to independently replicate their studies. Conflict of interest is another issue that has been raised by researchers (Abdalla and Abdalla, 2021; Young et al., 2022; Hagendorff and Meding, 2021). One notable case that exposed these concerns was that of Dr. Timnit Gebru, formerly part of Google's AI ethics team. Her dismissal highlighted the conflicts of interest that can arise when industry-aligned research agendas collide with ethical considerations. There is also evidence that the longer researchers remain in the industry, the more their high-impact work becomes privatized, limiting the dissemination of knowledge that could otherwise benefit the academic community (Jurowetzki et al., 2021).

While there are many concerns about potential threats to research integrity due to this influence, it is equally important to recognize the positive impact of Big Tech companies. First and foremost, their computational resources and data, previously inaccessible on such a scale, provide unprecedented opportunities for the discovery and development of novel Machine Learning (ML) algorithms. In addition, Big Tech–Academia collaboration may benefit scientific endeavors for reasons other than access to resources. Evans (2010) argues that the industry's relative lack of interest, in theory, could result in their academic partners being more likely to produce novel and theoretically unexpected experiments than academics working by themselves.

The matters mentioned above indicate that the influence of Big Tech (and more broadly the private sector, also referred to as *company* in our paper) on AI research needs to be carefully examined. Several such studies in this regard exist, including analysis of Big Tech's funding of academic AI researchers (Abdalla and Abdalla, 2021), changes in the proportion of Big Tech–affiliated papers at top conferences (Ahmed and Wahed, 2020), the flow of researchers from Academia to the private sector (Jurowetzki et al., 2021), the focus of the private sector on specific sub-fields of AI (Klinger et al., 2020), and the comparison of values encoded in the papers between the private sector and Academia (Birhane et al., 2021).

Unfortunately, approaches to this topic mentioned above are limited. The methods used so far to quantify Big Tech's dominance tend to oversimplify the concept of influence by perceiving it as the share of affiliations in papers regarding a specific topic. Methods are limited to tools such as topic modeling and keyword analysis. Current research also suffers from a lack of diversity in data sources, relying heavily on preprint servers such as arXiv or specific conferences to collect papers. In the previous studies mentioned above, the association of papers with affiliations was binary. That is, each paper was associated with either Big Tech (or the private sector in general) or Academia, depending on the first author or the proportion of affiliations in the paper. This limits conclusions that could be drawn from studies, as the Company–Academia collaboration in the area of AI is widespread.

All of the above limitations, combined with strong statements (e.g. researchers on this topic do not shy away from comparing Big Tech to Big Tobacco, Abdalla and Abdalla, 2021), could result in a limited, and possibly biased view of the matter, which could increase the tension between Big Tech and Academia, resulting in losses for science and society. Through our study, we aim to facilitate a more nuanced understanding of the dynamics between Big Tech and Academia and to highlight potential strengths and weaknesses within this relationship. We critically evaluate existing methodologies, advocate for more holistic measures of influence, and explore the implications of this dominance for the field of AI. We hope that this work will foster further discourse about the symbiosis between Academia and industry, and help to shape a more balanced AI research landscape.

In this study, along with a standard network analysis aimed at quantifying the influence of Big Tech and Academia on the AI papers citation network, we performed memetic analysis. A meme is understood here as a piece of an idea that is transmitted through culture. In the context of research papers, this

transmission occurs through citation. Using *meme score*, which quantifies the replicating power of a specific meme, we measure the spread of particular ideas in AI research. In addition, we quantify the probability of particular memes being replicated, conditioned on the affiliation of the authors of the paper containing the meme. This approach allows us to understand how the "spreading power" of specific ideas depends on affiliation groups (Big Tech, Academia, Companies, and mixed), which sheds light on which ideas Big Tech has more influence on spreading. Moreover, we investigated how papers with joint Big Tech–Academia affiliations differ from papers authored purely by authors affiliated with one category. In summary, we address the following research questions:

1. Do papers affiliated with both Big Tech and Academia differ in citation distributions from papers affiliated only with Big Tech or Academia? Do Big Tech affiliations differ from Company ones (i.e., industrial but not attributed to Big Tech)?
2. What ideas in AI research are the most contagious?
3. Does the contagiousness of a meme differ depending on the affiliation of paper authors?
4. What ideas are more contagious when discussed by Big Tech?

2. Related Work

2.1. Big tech influence over AI research

Several studies have attempted to quantify Big Tech's influence in AI research, using a variety of methods and data sources. Abdalla and Abdalla (2021) examine the funding provided by Big Tech to academic AI researchers. The study identifies a recurring pattern where private companies increase their support for academic institutions when their public image declines, often due to media incidents such as the Cambridge Analytica scandal. The authors compare this funding pattern to the tactics employed by the tobacco industry. The study reveals that 59% of papers published in top journals that address the ethical and societal implications of AI include at least one author with financial ties to a Big Tech company.

Ahmed and Wahed (2020) examine the changing dynamics of participation in major AI conferences following the rise of deep learning in 2012. The authors analyze 171,394 papers from 57 computer science conferences. They

find an increase in the participation of large technology companies, particularly since 2012. In addition, the paper uses term frequency analysis of abstracts to uncover distinct research areas among different organizations.

Jurowetzki et al. (2021) use bibliographic data to measure the flow of researchers from Academia to industry and examine the factors driving it. They find that 25% of the AI researchers moved to industry from institutions at the top 5 of the Nature Index. This observation suggests that industry tends to attract AI researchers from elite institutions, possibly reflecting a search for current and potential superstar talent or a narrow focus on high-prestige sources of talent.

Zhang et al. (2019) employs altmetrics to assess the influence of AI publications, revealing an increased public interest in AI research findings since 2011. However, the literature suggests that altmetrics may not be suited to measuring societal impact (Bornmann, 2014).

The recent article of Färber and Tampakis (2023) strikes to measure the relationship between authors' affiliations with the private sector and the article's popularity, measured by citations and attention score. They perform the keyword analysis with respect to the extent of association with the private sector. Their quantitative analysis shows the domination of the private sector in the AI research domain.

Krieger et al. (2021) note an upward trend in private sector publications in fundamental research journals as opposed to those centered on applied research. They also highlight a rise in collaborative publications with academic entities in contrast to solo publishing efforts.

Several studies have tried to analyze the content of AI papers to study the influence of Big Tech. These include topic modeling of the abstracts. Klinger et al. (2018) combine data from arXiv, GRID (Global Research Identifier) and MAG (Microsoft Academic Graph) to create a geocoded dataset of research activity in computer science disciplines. They identify deep learning papers using topic modeling. Finally, the authors measure the relatedness of computer science subjects based on their co-occurrence in arXiv papers. Klinger et al. (2020) analyze the field of study assigned to the papers based on arXiv. They find that companies focus more on applications of deep learning and on research advancing the computational infrastructure. AI techniques outside deep learning and broader AI applications, are of less interest to the private sector. Birhane et al. (2021) perform textual analysis in order to extract values encoded in papers. They analyze affiliations and funding sources in papers and find that the presence of Big Tech is increasing.

However, there is a significant gap in our understanding of Big Tech's influence. All of the above studies have at least one of the following limitations.

Firstly, none of the studies described above link the content of papers to citations in any way. Therefore, the resulting purely fraction-based analysis of content does not measure the spread of the ideas, but merely their prevalence. Quantifying both the spread and prevalence of specific ideas by using *meme score* allows us to find the most contagious ideas, called *memes*. The impact of affiliation groups on the contagiousness of each meme could be then modeled, resulting in a more fine-grained view of the influence of Big Tech, Academia, and other groups. Moreover, this approach allows for a comparison of the ideas on which contagiousness is most affected by Big Tech and Academia, which could provide insight into areas where Big Tech or Academia has the most influence.

Secondly, they simplify the notion of influence to the share of affiliations in papers regarding a specific topic. They use the share of papers affiliated with a particular group only (with the exception of Färber and Tampakis (2023)) that get published at prestigious conferences and in top journals as an indicator of the prevalence of Big Tech in AI research and interpret it as a proxy for influence measure. However, this approach does not take into account the number of the paper's citations, which is a more reliable proxy for the popularity of the paper and thus closer to measuring actual influence on the research field.

The third limitation is the quantity and diversity of data. The use of manual annotation limits the scope of the captured categories. Similarly, using only arXiv or specific conferences as the source of the papers limits the representativeness of the data.

2.1.1. Operationalization of the concept of Big Tech papers

Author affiliation is a key characteristic associated with the author rather than the research paper itself, although even a simple quantitative approach (i.e., the number of authors) has been proven to be connected to the impact exerted by an article (Sienkiewicz and Altmann, 2016). When examining the impact of Big Tech companies on scientific research, it becomes necessary to determine how to classify papers that are not exclusively authored by individuals affiliated with Academia or Big Tech. Existing approaches (Klinger et al., 2020; Ahmed and Wahed, 2020; Birhane et al., 2021) typically assume that any paper with at least one author affiliated with a Big Tech company

should be categorized as a Big Tech paper. Alternatively, they consider the affiliation of the first author as the defining criterion.

However, these approaches oversimplify a more complex situation. Grouping papers authored exclusively by researchers from the private sector, along with papers primarily written by academics, one of whom may have a dual affiliation, seems counter-intuitive. It has been observed that researchers associated with Big Tech companies benefit from access to computational resources, tools, and datasets that would otherwise be unavailable to them (Whittaker, 2021; Jurowetzki et al., 2021). Considering only the first author's affiliation would lead to overlooking these phenomena.

Another approach is to look separately at papers that result from collaborations between Big Tech companies and other institutions. Articles describing such collaborations tend to have different characteristics in terms of format and tone compared to the research papers mentioned above, and focus on the positive aspects of such collaborations (Popkin, 2019). In recent work, (Färber and Tampakis, 2023) separatly analyzed papers co-authored by Academia and company-affiliated researchers. To the best of our knowledge, no analyses have specifically examined the influence and topics of interest reflected in papers resulting from collaborations between Big Tech and academia.

Furthermore, we are not aware of any analysis that compares Big Tech-affiliated papers to those affiliated with the private sector in general. This limits the ability to assess whether found characteristics are specific to Big Tech, or are phenomena visible in all research originating from the private sector. In our study, we examine the differences between several affiliation groups: purely Big Tech, purely academic, private sector as a whole, and papers affiliated with both Big Tech/private sector and Academia.

2.2. Memetic analysis

Several measures can characterize the spread of ideas in science (Wagner et al., 2011; Xu et al., 2018) or in business research (Wu et al., 2017). A straightforward method to identify ideas within a set of documents is to examine their frequency. Techniques like tf-idf (term frequency–inverse document frequency) allow the discovery of potential keywords. Topic modeling methods, such as Latent Dirichlet Allocation (LDA) or BERTopic (Grootendorst, 2022), help to understand the distribution of ideas consisting of multiple phrases across a document set. Another option is to focus on topical clusters, which can be built from bibliographic coupling and co-citation

networks (Liu et al., 2017, 2019). One can use a citations cascade (Min et al., 2021) or chains of citations (della Briotta Parolo et al., 2020), which reveal a higher-order (i.e., reaching further than just the neighborhood of the node) temporal relationship between the papers.

One perspective on the ideas' propagation is looking at it through the lens of the evolutionary theory of science (Kantorovich, 2014). Central to this theory is the concept of *meme*, first introduced in a broader context by Dawkins (1976). A meme is considered a cultural analog to a biological gene, a unit of information that replicates itself, mutates, and undergoes selection in the evolutionary process. Analogous to genes in biology, memes are pieces of information transmitted between individuals within a given culture.

Kuhn et al. (2014) proposed an operationalization of the concept of a meme in the context of scientific knowledge by introducing the notion of a *meme score*. The meme score measures how much a given term is a meme by analyzing the citation network. It takes into account the term's frequency, the probability of its transmission through citation, and the probability of it emerging independently. In this operationalization of memes, a term exhibits more memetic quality, the higher the meme score of that term is. Since its proposal, the meme score has been used in several studies. It has been used to investigate how the gender of researchers influences their positions in the scientific community (Araújo and Fontainha, 2018), to identify scientific and technological trajectories of ideas in paper and patent networks (Sun and Ding, 2018), and to explore diffusion cascades of ideas (Mao et al., 2020).

3. Materials and methods

3.1. Meme score

The *meme score* was introduced by Kuhn et al. (2014) as a measure of the 'contagiousness' of meme.

The meme score for the given meme is calculated by multiplying the relative frequency of a term by its *sticking factor* and then dividing it by its *sparking factor*. The quotient of the sticking factor and the sparking factor is labeled as a propagation score. Both measures are briefly described below.

The sticking factor can be interpreted as the probability of the meme m being transmitted from one paper to the other paper. It corresponds to the probability that the meme appears in the abstract of the given paper, given that it appears in at least one of the abstracts of the papers cited by it. It is calculated by dividing the number of papers that contain the meme m and

cite at least one paper with this meme by the number of papers that cite at least one paper with this meme: $d_{m\to m}$ by $d_{\to m}$.

The sparking factor can be interpreted as the probability of the meme appearing without being present in any of the cited papers - how likely it is that the meme will be mentioned in the abstract, given that it does not appear in any of the cited papers' abstracts. It is calculated by dividing the number of papers with the given meme that do not cite any papers with this meme $d_{m\to\not m}$ by $d_{\to\not m}$ - all papers that do not cite any papers with the given meme.

The formula for the meme score can be written as follows:

$$M_m = \frac{N_m}{N}\frac{d_{m\to m}}{d_{\to m}}\backslash\frac{d_{m\to\not m}}{d_{\to\not m}} \tag{1}$$

where N_m is the number of papers with the meme m, N is the number of all memes in the dataset. $d_{m\to m}$ is the number of papers that contain the meme and cite at least one paper with the meme and $d_{\to m}$ is the number of papers that cite at least one paper with the meme. On the other hand $d_{m\to\not m}$ is the number of papers containing the meme but not citing a paper with the meme, and $d_{\to\not m}$ is the number of papers that do not cite a paper with the given meme.

From the three components of the meme score, it is mainly the sticking factor that measures the meme's contagiousness. The other factors control if the phrase is not merely used as a part of a natural language or if a phrase is popular enough. Due to this, the sticking factor on its own could serve as a weaker measure of contagiousness.

Let us underline that meme score can be interpreted as a ratio of the probability of replicating a phrase to the probability of the probability of its appearance without earlier presence in the cited papers. To select meaningful memes, we must select a threshold of the meme score defining which phrases are considered memes and which are not. This is done by selecting the point of maximum curvature from the value of the meme score as a function of its position in the phrase ranking ordered by the meme score (see Sec. 4.2 for practical implementation of referred method in this study). Such an approach is similar to the so-called "elbow method" (Ketchen and Shook, 1996) – a commonly used visual technique in assessing the optimal number of clusters in the k-means method.

3.2. Conditioned sticking factor

To measure how probable a meme is to spread under a given affiliation, we introduce a *conditioned sticking factor* – a modified version of the meme score introduced by Kuhn et al. A similar approach was proposed in the paper of Araújo and Fontainha (2018), where the authors attempted to condition the meme score on the main author's gender. In contrast, in our approach, we condition only the sticking factor on the authors' affiliations. The conditioned sticking factor is calculated as

$$\sigma_{m,a} = \frac{d_{m\to m,a}}{d_{\to m,a}}. \tag{2}$$

The conditioned sticking factor divides the number of papers $d_{m\to m,a}$ that have meme m and cite at least one paper with this meme and given affiliation a by the number of papers $d_{\to m,a}$ that cite at least one paper with given meme m and affiliation a. It can be interpreted as the probability of the meme transitioning during citation under the condition that the meme appeared in the paper with a given affiliation.

In contrast to the earlier approach of Araújo and Fontainha (2018), we have decided not to condition the sparking factor. The original purpose of the sparking factor is to control how often the word appears without any influence and limit the appearance of the most popular words that do not convey specific ideas. The sparking factor, conditioned analogously to the sticking factor, would correspond to the probability that the meme appears in a paper without being present in any cited papers with a given affiliation. However, the meme can appear in the other cited papers with different affiliations. In this way, the conditioned sparking factor would lose its original purpose of excluding phrases not conveying ideas without providing us with any specific additional information. For this reason, we decided to condition only the sticking factor. Finally, once $N_m \gg d_{m\to m}$, i.e., the meme is relatively common but rarely transmitted through the citation network, the meme score can be approximated by the sticking factor. Due to this, we focus primarily on the relationships between conditioned sticking factors for different groups. Figure 1 presents a toy example to explain how the sticking factor, sparking factor, meme score and conditioned sticking factor are calculated.

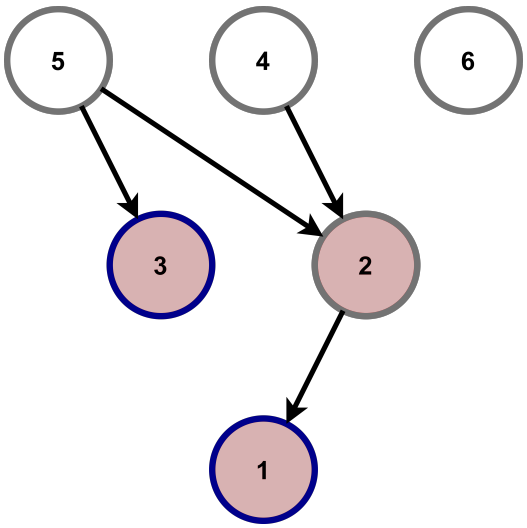

Figure 1: **Toy example of how to calculate the meme score.** Papers 1, 2 and 3 contain the meme (maroon background). Papers 1 and 3 have a given affiliation (dark blue frame) for which we calculate the conditioned sticking factor. The frequency of the meme is $\frac{3}{6}$. The unconditioned sticking factor is $\frac{1}{3}$, because of three papers that cite papers with the meme (2, 4 and 5), only one has the meme. The sparking factor is $\frac{2}{3}$, since the meme appears in two of three papers that do not cite papers with this meme – it appears in 1 and 3, but not in 6. The conditioned sticking factor is $\frac{1}{2}$, since meme appears in 1 out of 2 citations of the paper with the meme and given affiliation (2 replicates the meme from 1, 5 does not replicate the meme from 3).

3.3. Papers dataset and processing

As a main data source for this study, we used the S2ORC database Lo et al. (2020), which is a large corpus of 81.1 million English-language academic papers spanning many disciplines. The collection comes from various sources, such as in-proceedings of scientific conferences, well-established journals, as well as digital archives (e.g. arXiv). We gathered 557 681 articles from this source by searching for phrases presented in Appendix A) which occurred in the titles and abstracts of the papers. The aforementioned set of keywords is proposed in Liu et al. (2021). The authors of this paper conducted the study, attempting to find a representative set of keywords for the artificial intelligence domain while maintaining a reasonable balance between recall and precision.

Afterwards, we performed exploratory data analysis. We found out that 138 878 records had no DOI identifiers, which we used to match the papers against the OpenAlex database. As affiliation data provided by those links is crucial for our study, we decided to remove such papers from our dataset. Next, we discovered that 146 525 of the remaining records had no information about citations. Papers like this were useless for citation network analysis, so we decided to remove them from our dataset. Finally, we ended up with 166 455 records that had affiliations matched with OpenAlex.

The whole pipeline of memes extraction is shown in Fig. A.10. All used abstracts were preprocessed first – we lowercased all words and tokenized them. For the extraction of noun chunks, we used the `spaCy` library with the `en_core_web_md` model.

Papers' affiliation

First, we look at the category of the company as it is specified in the OpenAlex Database. OpenAlex provides information about the type of institution – it categorizes them into the following categories: company, education, government, facility, healthcare, nonprofit, and other (Table B.2). By analyzing these categories, we decided to merge education, facility, and government into Academia. We also analyzed the company category separately.

As Big Tech, we take a subset of technological companies that were among the top 60 companies by market capitalization in the years 2016-2019 (PwC, 2019, 2018, 2017, 2016). The following companies fall into this category: Apple, Google, Microsoft, Facebook, Tencent, Oracle, Intel, IBM, Cisco Systems, TSMC, SAP, Qualcomm, Amazon, Siemens, Alibaba Group, Nvidia, and Samsung. We aimed to construct a list that encompasses companies

previously labelled as Big Techs in media, such as the Big Five, FAANG or MAGA (Economist, 2018), while also being supported by a quantitative measure like market capitalization. To ensure the list's robustness and account for short-term market changes, we use a union of companies that appeared in the top 60 companies during the four years with the highest number of papers in our dataset. We opt to limit ourselves to the top 60 companies to include companies considered most significant to the field while excluding the lower end of the top 100 companies, where most technological companies appeared only once during the four years we consider.

Papers affiliated with Big Tech are 38% of the papers affiliated with the company category, but they account for 57% of their citations.

4. Results

4.1. Non-binary distinction between Big Tech and academy–affiliated papers

We provide substantial arguments, advocating for the less coarse distinction between Academia and Big Tech affiliations than the bisection often used in the literature. The most natural way to follow this proposition would be to use a continuous space by simply calculating for each paper the proportion of Big Tech affiliations to the total number of affiliations. However, as shown in Fig. 2, these values are unevenly distributed – the vast majority are simply academic papers (almost 98% of all papers in the examined dataset), around 0.5% belong only to the Big Tech class, and the remaining part is characterized by a mixed affiliation. The last group also suffers from obvious finite-size effects, i.e., due to a limited number of authors, fractions such as $\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$, and $\frac{3}{4}$ tend to occur more frequently than other values (see Fig. 2). To avoid the above obstacles and to be consistent with the observations presented in Fig. 2, we decided to use a *ternary* affiliation scheme, i.e., Academia–mixed Big Tech–Big Tech.

Let us first explore the implications of the ternary division using simple network metrics: incoming degree and PageRank (Brin and Page, 1998). Both these measures can be used to quantify the importance of a node (paper). In the case of the in-degree, we simply tally the number of citations a paper has received, while PageRank provides more nuanced information. Figure 3 shows distributions of in-degree (panel a) and PageRank (panel c), respectively, grouped according to the introduced ternary affiliation classification (denoted as "Academia", "Mixed" and "Big Tech"). In both cases, the Kruskal-Wallis test (Kruskal and Wallis, 1952) finds significant differences

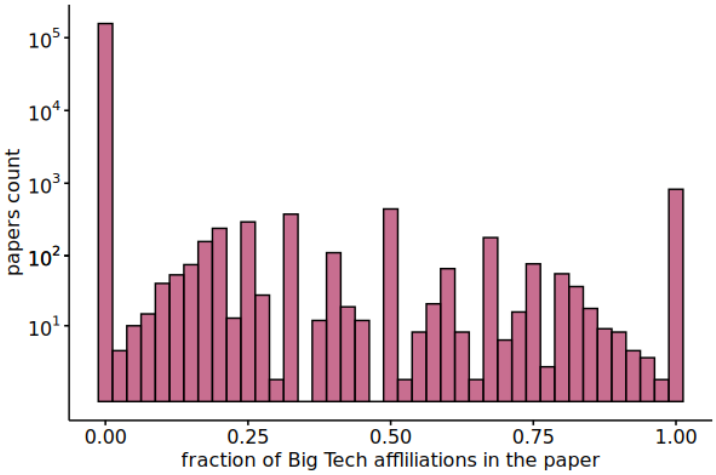

Figure 2: **Histogram (log scale) of the number of papers for different fractions of Big Tech affiliations**. The length of each bin is 0.025. Most papers are affiliated only with Academia, yet we can also witness a visible peak for purely Big Tech papers, which underlines the validity of this group. The plot also suggests that we should consider a third group of mixed affiliations, as the number of articles with Big Tech fraction is even bigger than that of Big Tech papers.

among all groups (p-value $< .001$), while post-hoc pairwise comparisons using Dunn's test (Dunn, 1964) with Holm-Bonferroni corrections (Holm, 1979) indicate that the described distinctions are maintained for all pairs of groups (if two or more categories are statistically indistinguishable, they are connected with a solid line on the far right part of the panel). To prove that the division into three categories does not lead to a loss of additional information available in the continuous approach, we divide the dataset into six categories: the first two and the last one are the same as in the previous division, while the remaining three describe cumulative quartiles without exclusive Big Tech affiliations (">25%", ">50%" and ">75%", i.e., ">25%" means that each node in this category has over 1/4 share of Big Tech affiliations but cannot be exclusively Big Tech). Indeed, also at this level, we obtain similar results for the Kruskal-Wallis test (significant differences with p-value $< .001$) as well as for pairwise comparisons, the only exception being the penultimate category (">75%"), which is indistinguishable from the pure Big Tech affiliations in the case of PageRank. All the intermediate categories ("Mixed", ">25%", ">50%" and ">75%") are indistinguishable from each other, which proves that performing further divisions in the "Mixed"

category does not bring any additional insights.

To examine the reasons for the clear differences between the elements of our ternary classification, we focus on the in-degree and PageRank probability distributions shown in Fig. 3b and Fig. 3d, respectively. Clearly, mixed and Big Tech groups differ from Academia by being characterized by higher values of probability density for larger values of in-degree or PageRank. On the other hand, the difference between mixed and Big Tech is not as pronounced. The likely main reason for the observed differences is revealed when we focus on the number of zero-degree nodes – the proportion of such nodes in Academia, mixed and Big Tech is 0.648, 0.538, and 0.613, respectively. Although the ratio of zero-degree nodes in the mixed category is visibly lower than in the other two categories, we confirm this observation by performing pairwise two-sample proportion tests with Holm-Bonferroni adjustment, which give p-values of $p < .0001$ for Academia–mixed Big Tech, $p = .044$ for Academia – Big Tech, and $p = .0004$ for mixed Big Tech– Big Tech. In effect, the network with removed zero-degree (isolated) nodes brings different statistical results marked in Fig. 3a and Fig. 3c by the dotted line on the far right – in this case we recover the binary division as only "Academia" category is distinguishable from any other.

Lastly, the decision to introduce the Big Tech category instead of looking at the OpenAlex company category was supported by the intuition that Big Techs differ significantly from other company actors and that extrapolating conclusions from all private companies to Big Tech would be unjustified. To test whether Big Tech are indeed different from the rest of the private companies, we compare the PageRanks of Big Tech–affiliated and company-affiliated papers. We perform the Kruskal-Wallis test between the PageRanks of Big Tech–affiliated papers and papers affiliated with companies that are not Big Tech. The test results confirm that the distribution of PageRanks of Big Tech-affiliated papers is indeed different from that of other companies (p-value < 0.05). Nevertheless, analyzing the ternary division from the company's point of view (instead of Big Tech) does not bring any additional insights, as can be seen in Fig. 4.

4.2. Most contagious ideas in AI research

Using the meme score measure, we were able to identify and analyze which ideas are the most prevalent in AI research in general, which is the first step in differentiating Big Tech from Academia.

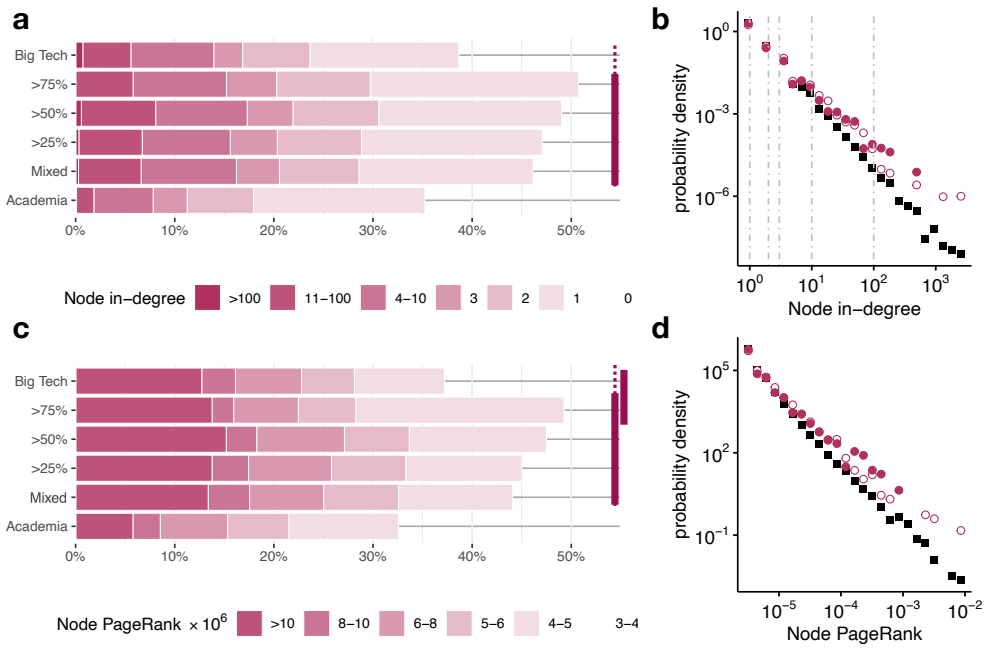

Figure 3: **Differences between Academia–mixed Big Tech–Big Tech as seen from the citation network perspective**. The rows are, respectively in-degrees of the nodes (panels a, b) and the PageRank values of the nodes (panels c, d). The left column shows the distribution plot, presenting the percentage of nodes with a given in-degree or PageRank value for ternary classification ("Academia", "Mixed", and "Big Tech") as well as cumulative quartiles (see text for details). Vertical solid lines on the far right of panels a and c connect statistically indistinguishable categories, while vertical dotted lines extend solid lines for the case of the network with removed isolated nodes. The right column presents probability density distributions associated with the ternary classification on panels a and b: filled squares represent Academia, empty circles – mixed Big Tech, and filled circles – Big Tech (node in-degree is increased by one to overcome log scale issues).

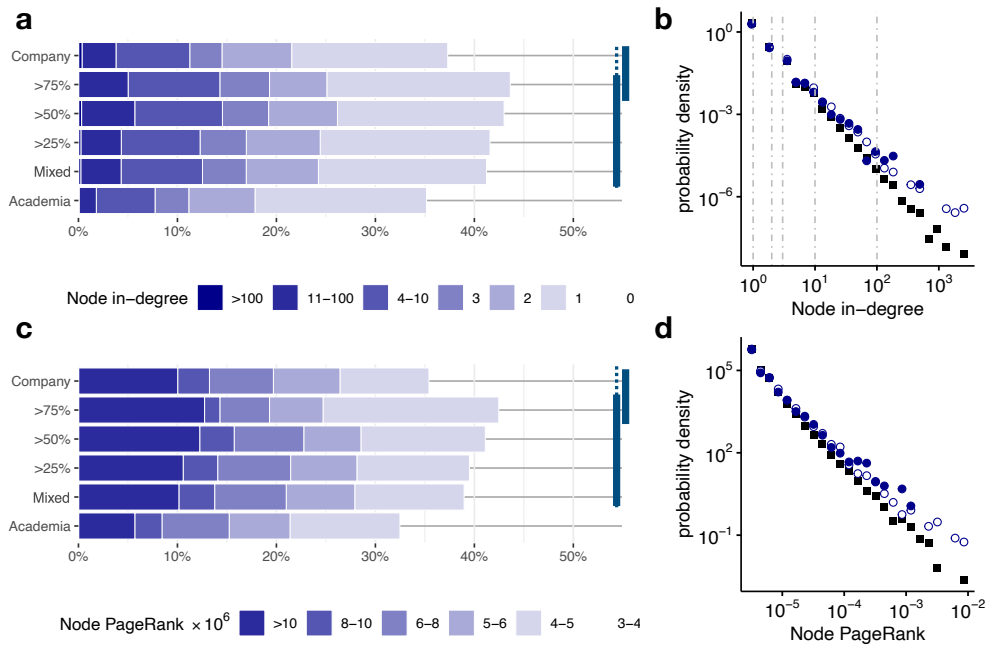

Figure 4: **Differences between Academia–mixed Company–Company as seen from the citation network perspective**. The rows are, respectively, in-degrees of the nodes (panels a, b) and the PageRank values of the nodes (panels c, d). The left column shows the distribution plot, presenting the percentage of nodes with a given in-degree or PageRank value for ternary classification ("Academia", "Mixed", and "Company") as well as cumulative quartiles (see text for details). Vertical solid lines on the far right of panels a and c connect statistically indistinguishable categories, while vertical dotted lines extend solid lines for the case of the network with removed isolated nodes. The right column presents probability density distributions associated with the ternary classification on panels a and b: filled squares represent Academia, empty circles – mixed Company, and filled circles – Company (node in-degree is increased by one to overcome log scale issues).

To perform this analysis, we calculated the meme scores and propagation scores and counted the occurrences for all memes present in our dataset, which resulted in over 1.7 million observations. Such a vast amount of data had to be filtered to get the most interesting results, which is why we decided to limit our study to phrases where the meme score was greater than 0, which resulted in over 60,000 meaningful observations.

Since our main goal is to deeply analyze the most spreading ideas, we limited the number of memes according to two criteria: the observed occurrences and their meme score values. The first criterion is based on the idea that the memes that have high meme scores but only occur once or twice are not really meaningful topics, as they are too niche to be considered contagious. The second condition is obvious since the higher the meme score, the more likely it is to spread. To choose a cutoff value, we analyzed the curves shown in Fig. 5. Finally, after limiting ourselves to phrases with a meme score above 0.25 and at least 20 occurrences, we ended up with 251 observations representing the top memes in our dataset. A manual annotation and inspection of the selected memes yields 80% precision, which is a comparable result to the one reported by Kuhn et al. 2014 (i.e., around 81.2% on 150 memes).

The set of chosen top memes was later annotated by the group of domain experts, which is briefly described in Appendix C, resulting in the assignment of one category per observation. The resulting categories are *ML algorithms*, *Medical terms*, *Security*, *Niche ML applications*, *ML Concept*, *Maths*, *Graphs*, *Computer Vision*, *Natural Language Processing*, *Data Related*, and *Other*. The annotations of the top memes allow us to analyze which high-level topics are popular among ML researchers, indicating the most promising research areas. According to Table 1, the most popular topic among scientists is the medical applications of ML. This category has the most memes and a third number of the memes occurrences in our dataset. It is not the only application, however, as we can see on the top memes list other important topics like security, and a whole category describing niche applications of ML solutions. However, this doesn't mean that today's research focuses only on ML applications, as the second-highest average meme score value, and the second sum of the occurrences belong to the topic describing the ML algorithms. Additionally, we were able to distinguish some particularly popular directions of AI development, such as Computer Vision or NLP. Finally, such an in-depth analysis proves the quality of our framework as the category Other, which includes unrelated topics, as well as meaningless

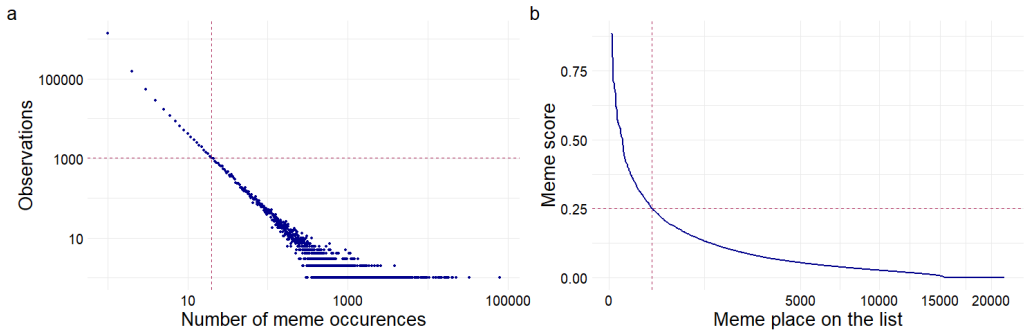

Figure 5: **Top memes selection by thresholds.** The left plot shows the number of remaining phrases, depending on the minimum number of meme occurrences. Since the dispersion of observations occurs around the number 20 on the x-axis, and 1000 on the y-axis (marked as dashed lines), we decided to use this point as a cutoff, which resulted in limiting ourselves to over 20 000 observations that have more than 20 appearances. The right plot shows the meme score value of the observation as a function of its position in the list sorted by the meme score. Since the major flattening occurs around the meme score value of 0.25, we decided to use this point as a cutoff, which resulted in limiting ourselves to 251 observations (marked as dashed lines).

memes, is responsible for only 20% of detected memes. The group has the biggest number of occurrences due to the memes such as 'paper', 'results', or 'art' (part of 'state of the art' phrase), which occur multiple times in the majority of the papers.

As we already know the big picture of the main topics in the scientist's interests, we can combine it with a more detailed analysis, and provide a wider perspective for our top memes group.

Figure 6 shows the characteristics of the top memes subset. As the meme score value depends on the number of memes occurrences as well as the propagation score we can easily imagine that the line that stands for the meme score equal to 0.25 is used as a cutoff value in this analysis. From the density of observations, especially in our group, we can grasp the tendency that the highest meme scores were assigned for the observations with a smaller number of occurrences. Additionally, the plot helps us detect the outliers aforementioned in the previous section, as they yield low propagation score, but also have a large number of occurrences. Thanks to the exemplary meme annotations we can also track some memes and analyze the reasons why are they considered important. For example, 'fairness' is considered important

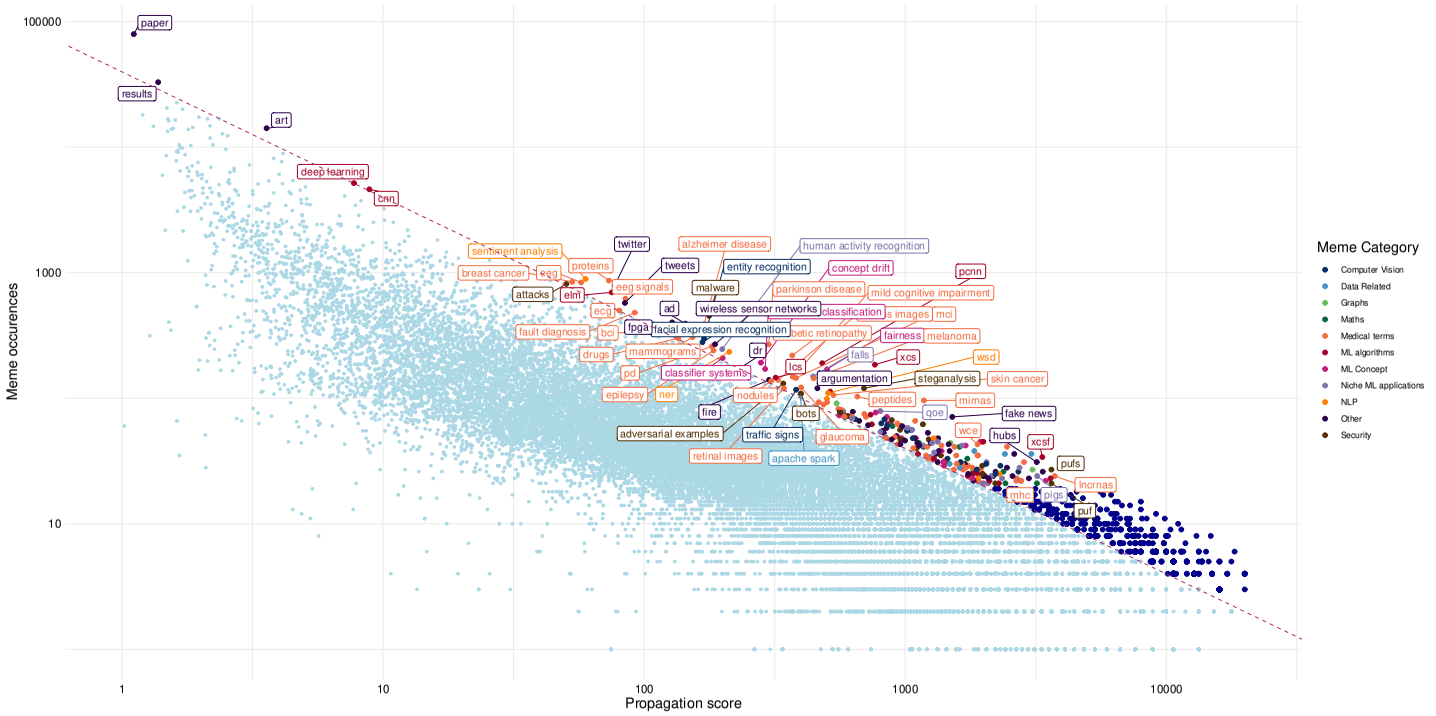

Figure 6: **Characteristics of top memes compared with other observations.** The observations on the plot are memes present in the whole dataset, and they are distinguished by the number of occurrences and propagation scores. Light-blue observations are memes with a meme score below 0.25, whereas the dark-blue ones have a score above 0.25, but the number of occurrences of these observations is below 20. The colorful dots represent the top memes with meme scores over 0.25, and more than 20 occurrences. These observations are additionally assigned to the meme categories described in the paragraph above. Exemplary observations were also labeled with their meme name.

Category	Number of memes	Mean meme score	Mean occurrences	Sum of occurrences
Medical term	70	0.339	141	9865
Other	50	0.347	2620	130993
Niche ML application	24	0.345	56	1336
Security	23	0.346	107	2454
ML Algorithm	20	0.380	581	11616
ML Concept	20	0.343	73	1456
Computer Vision	17	0.331	79	1336
Maths	9	0.348	35	312
Data Related	7	0.405	48	338
NLP	7	0.332	204	1430
Graphs	4	0.384	47	188

Table 1: **A statistical summary of the topics present in the top memes subset.** The table provides information about the number of memes existing within the described category. It also shows additional information describing the average meme score value, the average number of meme occurrences, and the sum of the total occurrences of memes from that topic.

mostly due to its high propagation score, whereas 'cnn' or 'deep learning' benefits from a vast number of occurrences.

To finalize the topic analysis, we decided to present all top memes in the form of a structured word cloud presented in Fig. 7. This way, we are able to analyze particular memes that build each of the predefined categories. We should particularly focus on the medical memes, as this topic is definitely the most important one, due to the abundance of different memes. The memes creating this topic (presented as orange), cover various medical terms. We can discover which disease treatments and studies greatly benefit from the usage of ML (ex. various cancers, epilepsy, diabetic retinopathy, Alzheimer's disease, malaria, etc.). We can also say that another major points of interest are the studies considering genomics (lncRNA, miRNAs, microRNAs, etc.), and that ML is widely used for various examinations (mammograms, eeg, ecg, wireless capsule endoscopy, etc.).

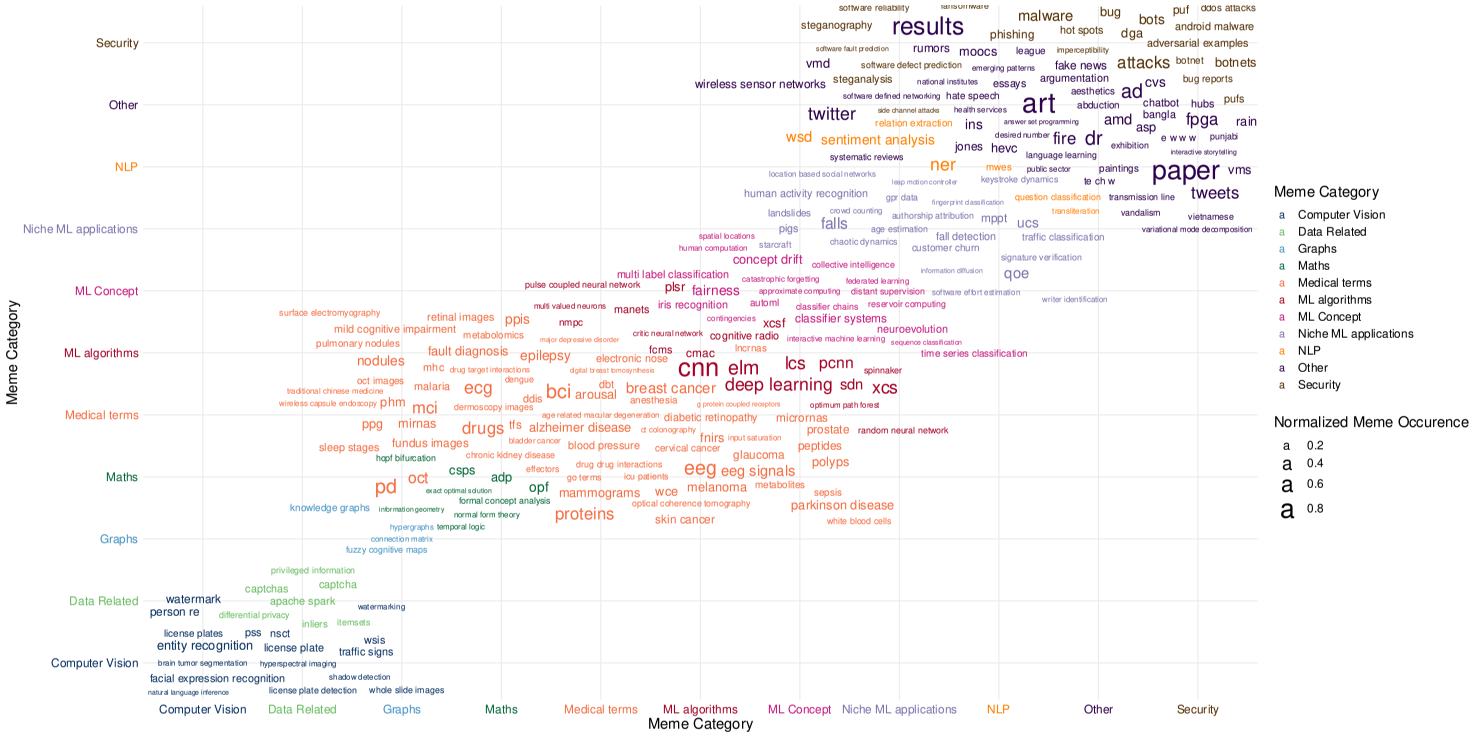

Figure 7: **A word cloud plot presenting top memes.** The size of the text represents the meme occurrence normalized to a [0, 1] range, whereas the color shows to which topic the meme belongs. The word cloud is organized in such a way that observations from the same topic are close to one another, as the X, and Y axes are discrete and consist of ordered categories.

4.3. Differences in contagiousness between Companies, Academia, and Big Tech

To assess whether contagiousness varies depending on author affiliation, we conducted a comparative analysis of conditioned sticking factors across different affiliation groups. A selection of the 15 highest-ranked memes in each affiliation (Academy, mixed Big Tech, Big Tech, mixed Company, and Company) ordered according to their conditioned sticking factor is shown in Fig. 8a. Although this limited picture suggests that *Academia* memes dominate, we need to underline that this is simply an illustrative excerpt from the whole distribution.

In order to perform a more systematical analysis, we compared all non-zero conditioned sticking factors across all affiliation groups using Kruskal-Wallis test. Results indicate a significant difference across groups in terms of the sticking factor, i.e. affiliation is a factor that creates distinctions among sticking factor values. Outcomes of post-hoc comparison performed using Dunn's test between categories presented in Fig. 8b pinpoint distinctions among the sticking factor distributions: strikingly, there are no statistical differences among Academia, Big Tech, and Company. On the other hand, such pairs as mixed Big Tech – Big Tech or mixed Company – Company can be distinguished. This is exactly opposite to the relations observed using network metrics for non-isolated nodes and shown in Figs. 3 and 4, where differences between mixed Big Tech – Big Tech and mixed Company – Company were negligible while one could clearly distinguish Academia, Big Tech, and Company categories.

The above analysis is conducted in absolute terms, i.e., the distributions can be biased by the fact that Academia affiliations are exceedingly more frequent than Big Tech or Company ones (cf Fig. 2). To overcome this issue we focused on memes that appear in at least two different affiliation categories, e.g., "random neural network" (present in mixed Company, Company, mixed Big Tech, and Big Tech) or "captcha" (present in Company, mixed Big Tech, and Big Tech) depicted in Fig. 8a. Following, we conducted Wilcoxon signed-rank tests (Wilcoxon, 1945), comparing sticking factors conditioned on specific groups using matched samples (the same memes in both groups). In case we were able to reject the null hypothesis (observations in both groups X and Y are exchangeable) in favor of the one-sided alternative hypothesis (differences $X - Y$ are stochastically larger than a distribution symmetric about zero) we noted it with a directed link in Fig. 8c, pointing from X to Y. The results revealed that memes affiliated with Big Tech

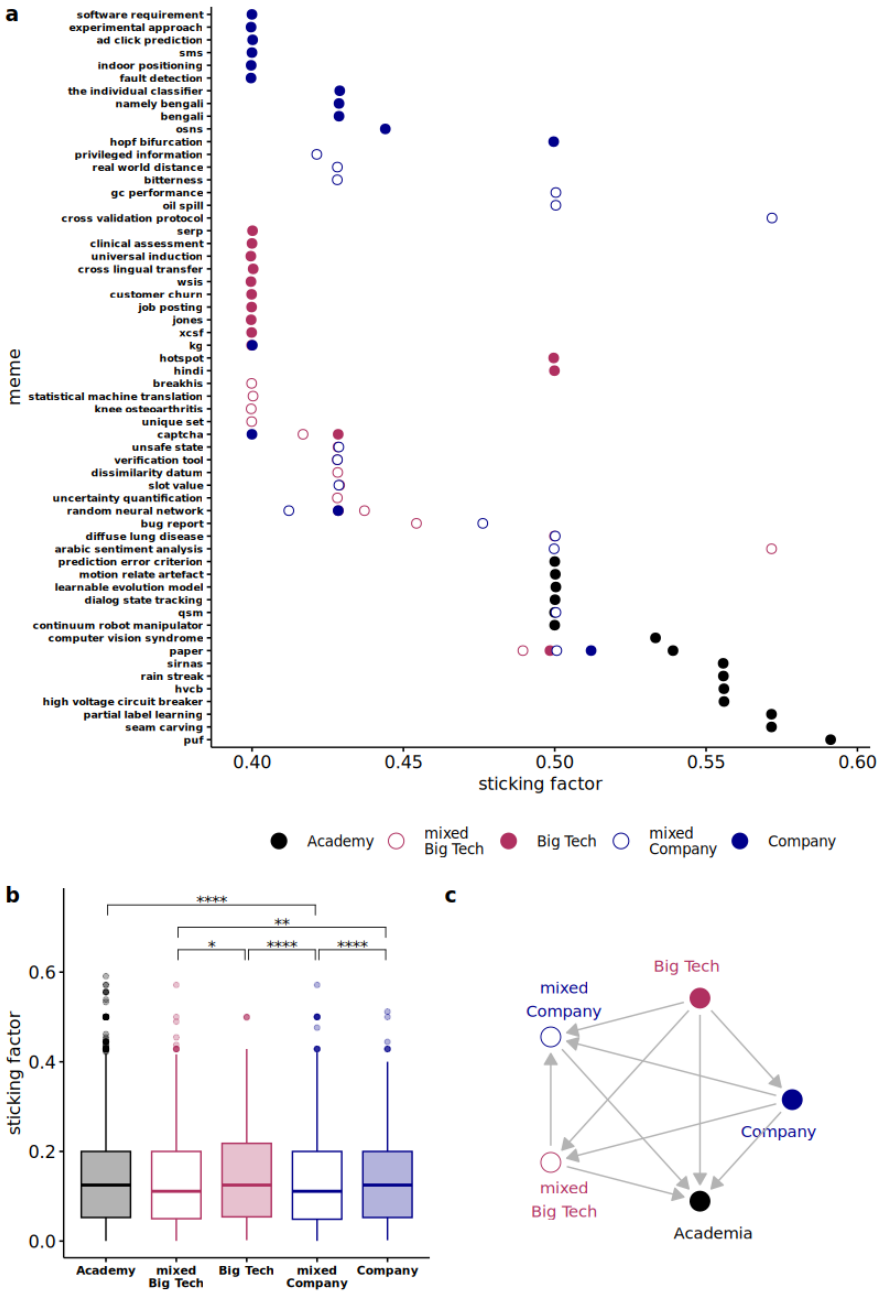

Figure 8: **Differences in contagiousness between Companies, Academia, and Big Tech.** (a) Selection of 15 memes with the highest sticking factor values in each category. (b) Boxplots of sticking factors conditioned on affiliation. The significance of post-hoc comparisons is indicated by *, **, ****, representing $.01 < p < .05$, $.001 < p < .01$ and $p < .0001$, respectively – insignificant comparisons are not shown. (c) Results of Wilcoxon tests for paired memes – an incoming arrow (pointing from category X to category Y) informs about statistically higher sticking factors of paired memes in X than in Y.

exhibited higher sticking factors compared to all other affiliation categories, with company-affiliated memes ranking second. Academic affiliation, on the other hand, was associated with a decrease in the sticking factor compared to other affiliations.

Following this, we explore how this difference depends on the meme category introduced in Table 1. Five hundred memes with the highest meme scores, excluding those not present at least once in papers from each affiliation group, were selected for this comparison. After removing duplicates and the least common categories of memes, the resulting list has 148 memes. The resulting comparison of conditioned sticking factors divided by meme category is presented in Fig. 9.

5. Discussion and conclusions

This work aimed to identify the most contagious ideas (memes) in AI research and assess to what extent the contagiousness of memes depends on the affiliation group of authors of the paper, and whether this contagiousness differs between pure and mixed academic or company affiliation. For this purpose, we analyzed AI papers' abstracts, affiliations, and citation networks. We employed two key metrics, namely the meme score and the sticking factor, to quantify the contagiousness of these memes. Additionally, we introduced the concept of conditioned sticking factor to assess the relationship between affiliations and contagiousness.

Our findings reveal several noteworthy insights. Firstly, we observed that papers affiliated with Big Tech tend to receive disproportionate citation counts in comparison to all companies. However, the most highly cited papers are those co-authored by individuals from both Big Tech and academic institutions, i.e., papers having mixed affiliations. Additionally, papers co-authored by individuals from both corporate and academic backgrounds exhibit scientometric distinctions from papers authored exclusively by Academia or corporate entities. This pattern holds true for both Big Tech and Company, thus challenging the binary distinctions often assumed in prior research.

The analysis of contagiousness indicates that on average, the contagiousness of Academia-authored memes does not differ from those authored by companies or Big Tech, although when taking into account only memes present in both affiliations, the same memes have on average higher contagiousness when mentioned by Big Tech than Academia. These facts together

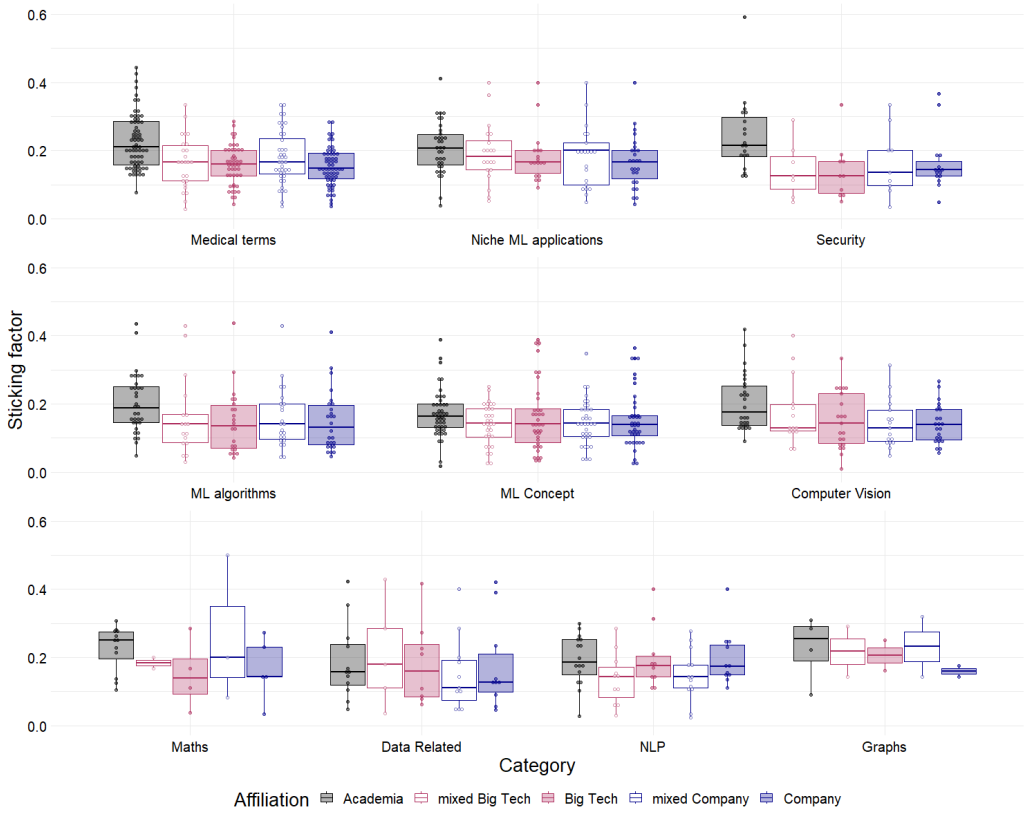

Figure 9: **Conditioned sticking factor across ML categories.** Box-plot depicts the distribution of the conditioned sticking factor of memes across different ML categories, stratified by the originating affiliation. The distribution patterns for most categories mirror the aggregate trends. In the 'Data Related' segment, memes generated by Big Tech exhibit a noticeably higher sticking factor, suggesting a dominant role in propagating novel content within this domain—likely a reflection of the substantial resources they possess. For 'Niche ML Applications', companies have a higher median sticking factor compared to their performance in other meme categories, implying their effectiveness in spreading memes that may complement their commercial strategies.

could be linked with the difference in the diversity of memes produced and propagated by those two groups. Academia creates far more memes than Big Tech or Companies (mostly due to the difference in overall papers authored), resulting in mean contagiousness being not significantly different than Big Tech or Company. When comparing common ideas of interest of those two groups, the difference in contagiousness becomes apparent. Those could be due to multiple factors, such as bigger resources available for Big Tech's researchers.

The impact of affiliation group over the contagiousness of memes could also be understood by applying the *framing theory*. Framing theory states that an issue can be perceived from various perspectives, each carrying implications for multiple values or considerations (Chong and Druckman, 2007). In the context of memetics, we can assert that each meme possesses the potential for diverse framings. These framings, tied to the affiliations of the authors, become integral components that propagate alongside the meme. Consequently, the contagiousness of a meme, when conditioned on a specific group, signifies the extent of that group's influence over the framing of the respective meme. In essence, the more contagious a meme becomes within a particular group, the more pronounced is that group's control over the framing associated with the meme. To assess this further, one could analyze the co-occurrence of memes, and ties of such memes "mixtures" to affiliations.

The analysis we have presented confronts the influence of Big Tech with that of academic institutions on the propagation of ideas between scientific papers. In doing so, we offer a unique perspective on tracing the evolution of ideas influenced by major technological companies and academic institutions. The conditioned sticking factor, introduced in this work, represents a novel tool that can extend beyond the confines of this study. It can be employed to analyze how various modalities, such as country of origin, journal, or institution, impact the spread of ideas in different contexts. In this way, possible outcomes could complement or challenge the results obtained in previous studies that tackled the problem of diffusion of ideas in a given scientific discipline (Hargreaves Heap and Parikh, 2005), or technology adoption under different (also institutional) constraints (Galang, 2014). In the same manner, if provided with additional data, the elaborated approach could be of use to track relations between organization characteristics and academia-industry interactions (see, e.g., Scandura and Iammarino, 2022). Last but not least, the conditioned meme score might be used to help measuring information overload (IOL, Hołyst et al. 2024) in science, which has lately become a

critical issue for researchers.

Further direction of research should also take into account the dynamics of memes contagiousness (i.e., equip the analysis with a dynamical point of view as it was done, e.g., by Kuhn et al. 2014) , investigating whether the impact of affiliation on contagiousness – potentially representing a gap between Academia, Companies, and Big Tech – is expanding or contracting would provide valuable insights. The relationship between Big Tech and Academia in AI could also be compared to parallel relations in non-AI computer science and in other disciplines, such as medicine, where the role of "Big Pharma" companies is considered significant (Giunta et al., 2016).

A limitation of this study, inherent in the original meme score calculating technique, is the operationalization of memes as simple noun chunks, overlooking the potential synonymity of phrases. Consequently, there is a possibility that high-spreading memes may be overlooked. Possible solutions could be techniques such as clustering embedded memes.

In conclusion, our results suggest that the prevalent notion of Big Tech's dominance over AI research is overly simplistic. We advocate for a more nuanced understanding of the roles played by Big Tech companies, corporations, and Academia in shaping the future of AI.

Acknowledgements

The research was funded by (POB Cybersecurity and Data Science) of Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. This work was also funded by the European Union under the Horizon Europe grant OMINO – Overcoming Multilevel INformation Overload (grant number 101086321, http://ominoproject.eu). Views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them. Computational part of this study was supported in part by the Poznań Supercomputing and Networking Center (grant number 607).

References

Mohamed Abdalla and Moustafa Abdalla. The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings*

of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, pages 287–297, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi:10.1145/3461702.3462563.

Nur Ahmed and Muntasir Wahed. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *arXiv:2010.15581 [cs]*, October 2020. URL http://arxiv.org/abs/2010.15581.

Tanya Araújo and Elsa Fontainha. Are scientific memes inherited differently from gendered authorship? *Scientometrics*, 117, 08 2018. doi:10.1007/s11192-018-2903-7.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research. *arXiv:2106.15590 [cs]*, June 2021. URL http://arxiv.org/abs/2106.15590.

Lutz Bornmann. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4):895–903, October 2014. ISSN 1751-1577. URL https://www.sciencedirect.com/science/article/pii/S1751157714000868.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi:https://doi.org/10.1016/S0169-7552(98)00110-X. Proceedings of the Seventh International World Wide Web Conference.

Dennis Chong and James N. Druckman. Framing Theory. *Annual Review of Political Science*, 10(1):103–126, June 2007. ISSN 1094-2939, 1545-1577. doi:10.1146/annurev.polisci.10.072805.103054.

Richard Dawkins. *The selfish gene*. Oxford University Press, New York, 1976. ISBN 978-0-19-857519-1.

Pietro della Briotta Parolo, Rainer Kujala, Kimmo Kaski, and Mikko Kivelä. Tracking the cumulative knowledge spreading in a comprehensive citation network. *Physical Review Research*, 2(1):013181, February 2020. doi:10.1103/PhysRevResearch.2.013181.

Olive Jean Dunn. Multiple comparisons using rank sums. *Technometrics*, 6 (3):241–252, 1964. doi:10.1080/00401706.1964.10490181.

Christoph Ebell, Ricardo Baeza-Yates, Richard Benjamins, Hengjin Cai, Mark Coeckelbergh, Tania Duarte, Merve Hickok, Aurelie Jacquet, Angela Kim, Joris Krijger, John MacIntyre, Piyush B. Madhamshettiwar, Lauren Maffeo, Jeanna Matthews, Larry Medsker, Peter Smith, Savannah Jennifer Thais, Savannah Thais, Savannah Thais, and Savannah Thais. Towards intellectual freedom in an AI ethics global community. *AI and Ethics*, 1 (2):131–138, 2021. doi:10.1007/s43681-021-00052-5.

The Economist. Move over faang, here comes maga – the tech giants are still in rude health. *The Economist*, August 2018. Retrieved 28 March, 2024.

James A. Evans. Industry induces academic science to know less about more. *American Journal of Sociology*, 116(2):389–452, 2010. doi:10.1086/653834.

Michael Färber and Lazaros Tampakis. Analyzing the impact of companies on AI research based on publications. *Scientometrics*, November 2023. ISSN 1588-2861. doi:10.1007/s11192-023-04867-3.

Roberto Martin N. Galang. Divergent diffusion: Understanding the interaction between institutions, firms, networks and knowledge in the international adoption of technology. *Journal of World Business*, 49(4):512–521, 2014. ISSN 1090-9516. doi:https://doi.org/10.1016/j.jwb.2013.12.005. URL https://www.sciencedirect.com/science/article/pii/S109095161300093X.

Anna Giunta, Filippo M. Pericoli, and Eleonora Pierucci. University–industry collaboration in the biopharmaceuticals: the italian case. *The Journal of Technology Transfer*, 41(4):818–840, Aug 2016. doi:10.1007/s10961-015-9402-2. URL https://doi.org/10.1007/s10961-015-9402-2.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

Thilo Hagendorff and Kristof Meding. Ethical considerations and statistical analysis of industry involvement in machine learning research. *Ai & Society*, pages 1–11, 2021. doi:10.1007/s00146-021-01284-z.

Shaun P. Hargreaves Heap and Ashok Parikh. The diffusion of ideas in the academy: A quantitative illustration from economics. *Research Policy*, 34(10):1619–1632, 2005. ISSN 0048-7333. doi:https://doi.org/10.1016/j.respol.2005.08.005. URL https://www.sciencedirect.com/science/article/pii/S0048733305001782.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL http://www.jstor.org/stable/4615733.

Janusz A. Hołyst, Philipp Mayr, Michael Thelwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N. Kenett, Denis Helic, Aljoša Rehar, Sebastijan R. Maček, Przemysław Kazienko, Tomasz Kajdanowicz, Przemysław Biecek, Boleslaw K. Szymanski, and Julian Sienkiewicz. Protect our environment from information overload. *Nature Human Behaviour*, 8(3):402–403, Mar 2024. ISSN 2397-3374. doi:10.1038/s41562-024-01833-8. URL https://doi.org/10.1038/s41562-024-01833-8.

Roman Jurowetzki, Daniel Hain, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain? *arXiv:2102.01648 [cs]*, February 2021. URL http://arxiv.org/abs/2102.01648.

Aharon Kantorovich. An evolutionary view of science: Imitation and memetics. *Social Science Information*, 53(3):363–373, 2014. ISSN 0539-0184. doi:10.1177/0539018414526325. Publisher: SAGE Publications Ltd.

David J. Ketchen and Christopher L. Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17(6):441–458, 1996. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0266%28199606%2917%3A6%3C441%3A%3AAID-SMJ819%3E3.0.CO%3B2-G.

Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. *arXiv:1808.06355 [cs, econ]*, August 2018. URL http://arxiv.org/abs/1808.06355.

Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. A narrowing of AI research. *Social Science Research Network*, 2020. doi:10.2139/ssrn.3698698.

Bastian Krieger, Maikel Pellens, Knut Blind, Sonia Gruber, and Torben Schubert. Are firms withdrawing from basic research? An analysis of firm-level publication behaviour in Germany. *Scientometrics*, 126(12):9677–9698, December 2021. ISSN 1588-2861. doi:10.1007/s11192-021-04147-y.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260): 583–621, 1952. doi:10.1080/01621459.1952.10483441.

Tobias Kuhn, Matjaž Perc, and Dirk Helbing. Inheritance Patterns in Citation Networks Reveal Scientific Memes. *Physical Review X*, 4(4):041036, November 2014. doi:10.1103/PhysRevX.4.041036.

Na Liu, Philip Shapira, and Xiaoxu Yue. Tracking developments in artificial intelligence research: constructing and applying a new search strategy. *Scientometrics*, 126(4):3153–3192, April 2021. ISSN 1588-2861. doi:10.1007/s11192-021-03868-4.

Wenyuan Liu, Andrea Nanetti, and Siew Ann Cheong. Knowledge evolution in physics research: An analysis of bibliographic coupling networks. *PLOS ONE*, 12(9):e0184821, 2017. ISSN 1932-6203. doi:10.1371/journal.pone.0184821.

Wenyuan Liu, Stanisław Saganowski, Przemysław Kazienko, and Siew Ann Cheong. Predicting the Evolution of Physics Research from a Complex Network Perspective. *Entropy*, 21(12):1152, December 2019. ISSN 1099-4300. doi:10.3390/e21121152.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL `https://api.semanticscholar.org/CorpusID:215416146`.

Jin Mao, Zhentao Liang, Yujie Cao, and Gang Li. Quantifying cross-disciplinary knowledge flow from the perspective of content: Introducing an approach based on knowledge memes. *Journal of Informetrics*, 14(4): 101092, 2020. ISSN 17511577. doi:10.1016/j.joi.2020.101092.

Chao Min, Qingyu Chen, Erjia Yan, Yi Bu, and Jianjun Sun. Citation cascade and the evolution of topic relevance. *Journal of the Association for Information Science and Technology*, 72(1):110–127, 2021. ISSN 2330-1643. doi:10.1002/asi.24370.

Gabriel Popkin. How scientists can team up with big tech. *Nature*, 565 (7741):665–667, January 2019. doi:10.1038/d41586-019-00290-y.

PwC. Global top 100 companies by market capitalisation 20176. Technical report, 2016. [Accessed 28-03-2024].

PwC. Global top 100 companies by market capitalisation 2017. Technical report, 2017. [Accessed 28-03-2024].

PwC. Global top 100 companies by market capitalisation 2018. Technical report, 2018. [Accessed 28-03-2024].

PwC. Global top 100 companies by market capitalisation 2019. Technical report, 2019. [Accessed 28-03-2024].

Alessandra Scandura and Simona Iammarino. Academic engagement with industry: the role of research quality and experience. *The Journal of Technology Transfer*, 47(4):1000–1036, August 2022. ISSN 1573-7047. doi:10.1007/s10961-021-09867-0. URL https://doi.org/10.1007/s10961-021-09867-0.

Julian Sienkiewicz and Eduardo G. Altmann. Impact of lexical and sentiment factors on the popularity of scientific papers. *Royal Society Open Science*, 3(6):160140, 2016. doi:10.1098/rsos.160140.

Xiaoling Sun and Kun Ding. Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents. *Scientometrics*, 116(3):1735–1748, 2018. ISSN 0138-9130, 1588-2861. doi:10.1007/s11192-018-2836-1.

Caroline S. Wagner, J. David Roessner, Kamau Bobb, Julie Thompson Klein, Kevin W. Boyack, Joann Keyton, Ismael Rafols, and Katy Börner. Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of Informetrics*, 5(1):14–26, 2011. ISSN 1751-1577. doi:https://doi.org/10.1016/j.joi.2010.06.004. URL https://www.sciencedirect.com/science/article/pii/S1751157710000581.

Meredith Whittaker. The steep cost of capture. *Interactions*, 28(6):50–55, November 2021. ISSN 1072-5520, 1558-3449. doi:10.1145/3488666.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi:https://doi.org/10.2307/3001968.

Chaojiang Wu, Chelsey Hill, and Erjia Yan. Disciplinary knowledge diffusion in business research. *Journal of Informetrics*, 11(2):655–668, 2017. ISSN 1751-1577. doi:https://doi.org/10.1016/j.joi.2017.04.005. URL https://www.sciencedirect.com/science/article/pii/S1751157716303455.

Jian Xu, Yi Bu, Ying Ding, Sinan Yang, Hongli Zhang, Chen Yu, and Lin Sun. Understanding the formation of interdisciplinary research from the perspective of keyword evolution: a case study on joint attention. *Scientometrics*, 117(2):973–995, Nov 2018. doi:10.1007/s11192-018-2897-1. URL https://doi.org/10.1007/s11192-018-2897-1.

Meg Young, Michael Katell, and P.M. Krafft. Confronting power and corporate capture at the FAccT conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1375–1386. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533194.

Xi Zhang, Xianhai Wang, Hongke Zhao, Patricia Ordóñez de Pablos, Yongqiang Sun, and Hui Xiong. An effectiveness analysis of altmetrics indices for different levels of artificial intelligence publications. *Scientometrics*, 119(3):1311–1344, June 2019. ISSN 1588-2861. doi:10.1007/s11192-019-03088-x.

Appendix A. Data gathering and processing

Following Liu et al. (2021) we employed simple search strategy for AI papers. We searched for following words from titles and abstracts: *artificial intelligence, machine learning, classifier, neural network, deep learning, data science, nlp, machine-learning, computer vision.*

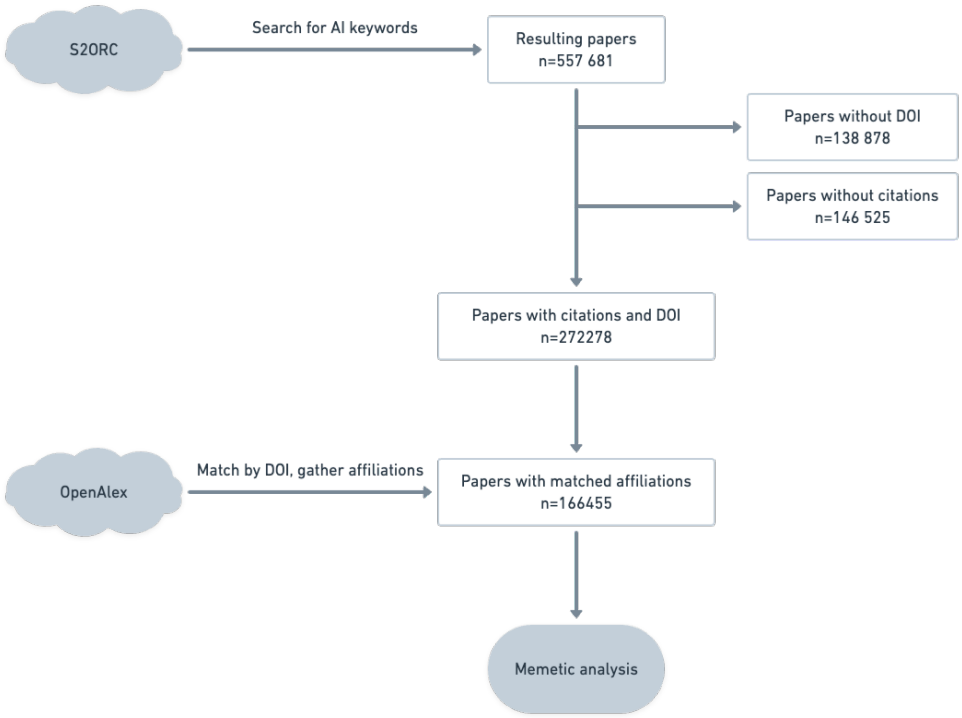

Figure A.10: Diagram of pipeline used for gathering and processing papers.

Appendix B. OpenAlex categories of affiliations treated as Academia

Category	Most important institutions
education	Tsinghua University
	Shanghai Jiao Tong University
	Carnegie Mellon University
	Stanford University
	Massachusetts Institute of Technology
government	Chinese Academy of Sciences
	French National Centre for Scientific Research
	National Institute of Health
	French Institute for Research in Computer Science and Automation
	Commonwealth Scientific and Industrial Research Organisation
nonprofit	Max Planck Society
	Electric Power Research Institute
	German Research Centre for Artificial Intelligence
	Instituto de Salud Carlos III
	SRI International
facility	Ecole Polytechnique Federale de Lausanne
	National Institute of Informatics
	Electronics and Telecomunications Research Institute
	United States Air Force Research Laboratory
	Italian Institute of Technology
healthcare	Mayo Clinic
	Boston Children's Hospital
	Brigham and Women's Hospital
	The University of Texas Southwestern Medical Centre
	Vanderbilt University Medical Centre

Table B.2: Five institutions with the biggest number of publications from each OpenAlex category included by us in the Academia category. Due to overlap in education, government and facility categories we decided to merge them into Academia.

Appendix C. Annotations

The annotation process was conducted for the top memes, defined as the ones that have a meme score value over 0.25, and more than 20 occurrences,

by the team of domain experts. The pipeline for the creation of the annotations consisted of several steps, described below.

1. Extraction of basic parameters: meme score, meme occurrences, and meme phrases.
2. Providing a short description of more complex, and not obvious memes. The research was focused on providing the descriptions connected to ML, as we know, that used dataset comes from this domain. Providing the titles of ML papers found connected to researched memes.
3. Providing the flags for duplicate or very similar memes. Most of them were the abbreviations and their elaborations.
4. Creation of first frequently repeating topics (ML algorithms, Medical terms, Security, Niche ML application, ML concept, and Other), based on the annotation process.
5. Iterative extraction of another topic from the Other group (Maths, Graphs, Computer Vision, NLP, Data Related).
6. Revising whole annotations and re-assignment to proper topics.