‘I SEE SOMETHING YOU DON’T SEE’. A COMPUTATIONAL ANALYSIS OF THE DIGITAL SERVICES ACT AND THE DIGITAL MARKETS ACT

Fabiana Di Porto,[*] *Tatjana Grote,*[**] *Gabriele Volpi*[***] *and Riccardo Invernizzi*[****]

ABSTRACT

The rise of digital platforms has been accompanied by increasing unease among scholars and policymakers about certain anti-competitive structures and practices. As a result, ensuring fair competition on digital services markets has been high up on the agendas of both US and EU lawmakers.

In its latest proposals, the Digital Markets Act and Digital Services Act (DMA and DSA), the European Commission puts forward several new obligations for online intermediaries, especially large online platforms and ‘gatekeepers’. Both are expected to serve as a blueprint for regulation in the US, where lawmakers have also been investigating competition on digital platforms.

This paper investigates whether key concepts of competition law on digital markets are used in the same way by different stakeholders. Leveraging the power of computational text analysis, we find significant differences in the employment of terms like ‘gatekeepers’, ‘self-preferencing’, ‘collusion’ and others in the position papers of the consultation process that informed the drafting of the two latest Commission proposals. Added to that, sentiment analysis shows that in some cases these differences also come with dissimilar attitudes. While this may not be surprising for new concepts such as gatekeepers, the same is not for others, like ‘self-regulatory’,

[*] Professor of Law and technology at the University of Salento, Italy (fdiporto@gmail.com). This paper is part of a broader interdisciplinary research project on Algorithmic disclosure regulation, funded by a three-year PRIN research grant awarded by the Italian Ministry of Research (no of grant 2017BAPSXF, www.lawandtechnology.it). We are thankful to Prof. Julian Nyarko from Stanford University for providing some of the data we used in the empirical part and to a number of other people for insightful discussions: Dr Roberto Marseglia, Prof. Daniela Piana, Giuseppe Italiano. All mistakes remain our own.

[**] Tatjana Grote is BA in International Relations from the Technical University of Dresden and a member of the PRIN research group on Algorithmic disclosure regulation.

[***] Gabriele Volpi, MSc Theoretical physics from the University of Pavia, Manager at DAAT consulting and Member of the PRIN Algorithmic Disclosure project (www.lawandtechnology.it).

[****] Riccardo Invernizzi, BSc in Mathematics (with honors) from the University of Pavia, junior analyst at DAAT consulting and Member of the PRIN Algorithmic Disclosure project (www.lawandtechnology.it).

which not only is used differently by stakeholders, but is also viewed more favorably by medium and big companies/organizations than by small ones. We conclude by sketching out how different computational text analysis tools (like e.g. topic modeling, semantic analysis and text similarity), could be combined to provide many helpful insights for both rulemakers and the legal scholarship.

Keywords: Digital Services Act, Digital Markets Act, Big Tech, Antitrust, Computational Analysis, Machine Learning, Competition, Gatekeepers, Remedies, Transparency duties

JEL: K3, K4, O3

Table of Contents

INTRODUCTION

"It is complex, but we are looking forward to it". This is the closing remark of EU Executive Vice-President Margarete Vestager[1] statement when introducing the latest Commission proposals on digital platforms: the Digital Markets Act and the Digital Services Act.[2]

Without a doubt managing competition on digital markets is not a simple endeavor. This is also reflected by the scholarly debate on new rules for digital markets, which has been burgeoning for a while.[3] In December 2020, the Commission published what everyone involved in these debates had been looking forward to: a twin-proposal suggesting many new rules for 'online platforms',[4] 'very large online platforms', and 'gatekeepers'.

These rules are not only relevant to EU businesses and consumers, but to businesses around the world, especially well-known US tech companies.[5] In fact, many expect that the DMA and DSA might serve as a 'blueprint for regulation across the globe', similar to the GDPR which has, *inter alia*,

[1] European Commission, Statement by Executive Vice-President Vestager on the Commission proposal on new rules for digital platforms, 15 December 2020, https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_20_2450 (accessed 15 February 2021).

[2] European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (COM(2020)825), 15 December 2020 [hereinafter Digital Services Act, DSA]; European Commission, Proposal for a Regulation of the European Parliament and the Council on contestable and fair markets in the digital sector (Digital Markets Act) (COM(2020) 842), 15 December 2020 [hereinafter Digital Markets Act, DMA].

[3] See, e.g., Ibáñez Colomo, P. (2020). Whatever Happened to the 'More Economics-Based Approach'?, Journal of European Competition Law & Practice 11, 9, 473–74, (discussing the shift from the so called 'more economic approach' to the growing demand for ex ante intervention against big digital platforms in the European legal community). For challenges related to competition law, see e.g. Ezrachi, A. & Stucke, M. (2016). *Virtual competition: the promise and perils of the algorithm-driven economy* (Cambridge: Harvard University Press), and Marsden, P., & Podszun, R. (2020). Restoring Balance to Digital Competition – Sensible Rules, Effective Enforcement, Konrad-Adenauer-Stiftung, 1-87. On consumer protection and its relation to data protection and competition law, see Kerber, W. (2016). Digital markets, data, and privacy: competition law, consumer law and data protection, *Journal of Intellectual Property Law & Practice*. 11(11), 856-866.

[4] There is no perfect overlapping between the legal definitions of 'platforms' in the DSA and DMA. In the DSA, the widest concept is that of online '*intermediary service*', which covers all services within the scope of Art. 1(3), including 'online platforms' (providing hosting services) (Art. 2(1)h)). Conversely, in the DMA, the widest concept is that of '*core online platform*' (Art. 2(2)), which covers 'online intermediation services' (inclusive of application stores), together with other services (like e.g. cloud computing services, social networking sites, videosharing platforms, search engines, operating systems, advertising services).

[5] See Art 1(3) DSA, Art 1(2) DMA.

influenced the 2018 California Privacy act.[6]

Given that US lawmakers, agencies and state attorneys general have been investigating competition on digital platforms recently,[7] the newly proposed EU rules might have some visible repercussions on the US legal landscape. To name just a few: if the acts are adopted, 'very large online platforms' (Art 25 DSA)[8] will need to *provide meaningful information* on how they manage (Art 23 DSA) and *rank* their content (Art 29 DSA).[9] The DMA complements these transparency-focused norms with specific do's and don'ts for platforms designated as 'gatekeepers' (Art 3 DMA):[10] Art 6(d) DMA prohibits self-preferencing; Art 5(e) and (f) DMA oblige gatekeepers to enable multi-homing and interoperability.[11]

As can be seen from these examples, terms and concepts like 'gatekeepers', 'self-preferencing', or 'interoperability' play a central role in designing new rules for online platforms. Inherently, this comes with the

---

[6] Jaursch, A. B. and J. (2020). How the EU plans to rewrite the rules for the internet. Brookings. https://www.brookings.edu/techstream/how-the-eu-plans-to-rewrite-the-rules-for-the-internet/ (accessed 5 February 2021); Voss, W. G., & Houser, K. A. (2019). Personal Data and the GDPR: Providing a Competitive Advantage for U.S. Companies. *American Business Law Journal* 56(2), 287.

[7] *See* US House of Representatives, Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary, *Investigation of Competition in Digital Markets* (2020), https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf. *See* also FTC v. Facebook, Inc., Compl. No. 191 0134 (Dist. Ct. D.C. 2020), 9 Dec. 2020; State of New York et al v. Facebook, Inc., Compl. No. 1:2020cv03589 (Dist. Ct. D.C. 2020), 9 Dec. 2020; U.S. v. Google LLC, Compl., No. 1:20-cv-03010 (Dist. Ct. D.C. 2020), 20 Oct. 2020. (accessed 5 February 2021).

[8] Some online platforms are subject to additional obligations: these are the 'very large online platforms', defined as those with at least 45 million monthly active users in the EU (Art. 25, DSA).

[9] Art. 29 DSA requires to disclose the main parameters used in recommender systems so to prevent self-preferencing practices in rankings.

[10] In the DMA, some providers of core online services may be designated as 'gatekeepers' if a three-criteria test is met. The latter is based on two different mechanisms: the default one is based on a turnover plus user thresholds and establishes a rebuttable presumption (Art. 3(2) DMA). If the thresholds are not met, the provider of a 'core platform services' can still be identified as a gatekeeper (Art. 3(6) DMA) after a market investigation by the Commission (Art. 15 DMA). Hence, not every very large platform is a gatekeeper, but it is likely that every gatekeeper will also be a very large online platform (see Art 3(2)b DMA).

[11] Multi-homing is the possibility to use the services of more than one platform simultaneously. I Interoperability is both the compatibility of protocols (protocol interoperability) and the possibility to access data in real-time for both the data subject and entities acting on the latter's behest (data interoperability). Crémer J., Y.-A. de Montjoye, H. Schweitzer (2019). Competition policy for the digital era, European Commission Report, https://data.europa.eu/doi/10.2763/407537 (accessed 14 February 2021) [hereinafter Crémer Report], 83-84.

challenge of defining these new concepts to make them legally operable.[12] Furthermore, there might be some unclarity on how exactly some more 'traditional' qualifiers, e.g., 'dominant' position or 'anti-competitive' conduct, apply to digital services.[13]

Consequently, a large part of the debate on new rules for digital markets prior to the publication of the EC twin-proposal has centered around how to define or understand certain essential concepts.[14]

This was well-reflected in the EC's public consultations, which allowed stakeholders to submit their statements on the Commission's initial impact assessments on the matter between 2 June and 8 September 2020.[15] These feedback documents are a valuable source to understand how those affected by new rules – ranging from individuals to platform giants – perceive the problems addressed and solutions proposed by the EC.

However, in light of the relative novelty of many terms and issues surrounding digital competition, one might ask whether simply reading these feedback documents suffices to actually develop a precise understanding of the stakeholders' take on certain critical questions. And therefore, whether the competitive issues at stake are adequately tackled by the proposed norms.

---

[12] DMA proposal, 7-8.

[13] On how to measure market power, see de Montjoye, Y.-A., Schweitzer, H., Crémer, J. (2019). Competition policy for the digital era, European Commission Report, https://data.europa.eu/doi/10.2763/407537 (accessed 14 February 2021) [hereinafter Crémer Report, 48-50.

[14] For instance, there has been a vivid debate on the scope of application of the Digital Services and the Digital Markets Act, in particular. For a comment on how to best set the criteria for application of the DMA, how to designate 'gatekeepers', see Geradin, D. (2020). The Digital Markets Act: How should ex ante rules look like? The Platform Law Blog. https://theplatformlaw.blog/2020/10/23/the-digital-markets-act-how-should-ex-ante-rules-look-like/ (accessed 10 May 2021). For a discussion of the relation between dominance and gatekeeping power, see Käseberg, T. (2020). 'Antitrust 2.0 – Governance of oversight over digital gatekeepers'. Kluwer Competition Law Blog. http://competitionlawblog.kluwercompetitionlaw.com/2020/12/14/antitrust-2-0-governance-of-oversight-over-digital-gatekeepers/ (accessed 10 May 2021).

[15] The DSA and DMA result from three different inception impact assessment documents: the **Digital Services Act,** Deepening the Internal Market and clarifying responsibilities for digital services, Inception impact assessment, Ares(2020)2877686, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services_en; the **Ex ante regulatory instrument for large online platform**s with significant network effects acting as gate-keepers in the European Union's internal market, Inception impact assessment, Ares(2020)2877647, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers; and the **New Competition Tool**, Inception impact assessment, Ares(2020)2877634, 4 June 2020, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12416-New-competition-tool (all accessed 31 March 2021).

Consider the following example: if all stakeholders demand new rules for 'gatekeepers' and 'very large online platforms' which explicitly prohibit 'self-preferencing' practices, how can we know they are having the same platforms and practices in mind? While some might think of gatekeepers as dominant platforms in the sense of Art 102 TFEU, others might set a much lower threshold. And while some might think that self-preferencing leads to foreclosure, others might view it as necessary to penetrate the market.

Stakeholders might concur at first sight, but some more subtle differences could be covered up by the use of terms which are still partially nebulous, either due to their novelty or because of the need to apply them to the digital sphere. Some of these differences might be discernible if all feedback documents are meticulously analyzed. This is something the Commission does mainly by means of qualitative analysis. However, language always has an element of subconsciousness, which often escapes the human eye, but can be traced with the help of algorithmic techniques, such as Natural Language Processing.

In this paper, we leverage the power of computational tools to analyze the debate preceding the publication of the DSA and DMA proposals. More specifically, we use supervised and unsupervised Machine Learning to address the following questions: (1) Does the use and understanding of terms related to competition in digital markets differ across different groups of stakeholders? (2) What are stakeholders' attitudes towards certain contentions terms?

On top of these substantiative questions, we further discuss (3) to what extent computational tools can help to automate and enrich the analysis of documents that are used as inputs to the rulemaking process. If this is the case, then such approaches should be recognized as a valuable addition to the 'manual' analysis the Commission currently undertakes.

This article is structured as follows. The following section sets the larger context by outlining some novel competition challenges posed by digital markets that led to the proposed ex ante regulatory response set out in the DSA and DMA (Ch. 1). We then present our methodology and show that (Ch. 2): first, similar opinions are expressed by groups usually belonging to different clusters (i.e. medium and big organizations); and second, groups of stakeholders use central terms of the DSA and DMA in different ways. Lastly, we conclude by suggesting what this evidence should tell about the two proposals and, on a more overarching tone, how computational methods could support EU targeting rules, although very cautiously.

I. COMPETITION IN THE DIGITAL ERA AND THE PROPOSED REGULATORY RESPONSE IN THE DSA AND DMA

There are manifold challenges arising from digital markets that standard competition rules seem not to tackle adequately. This section will give a snapshot of those inadequacies that were most debated, highlighting what ex ante rules received the greatest support as a consequence. We will then have a closer look at the DSA and DMA proposals, to see in what way such debate and the perceived needs translated into actual norms.

A. Why We Need the DSA and DMA

Consumers benefit in many ways from the impressive development of digital markets.[16] However, lawmakers and scholars alike have been emphasizing that digital markets show certain characteristics which are likely to favor highly concentrated markets. First, many business models in the digital realm are characterized by strong returns to scale.[17] Second, incumbents in online markets are particularly hard to dislodge due to substantial network effects.[18] Third, due to the data dependency of many online services, established players might hold a competitive edge over small contestants by leveraging the power of the large amounts of data they accumulate.[19]

While the extent of these advantages could be limited by multi-homing and interoperability,[20] there still is a very real chance that certain platforms accumulate some kind of 'gatekeeping' power and impose the prices, conditions, and level of transparency they deem appropriate, for their own convenience. In its investigation report on competition in digital markets, the

[16] See Recital 1 DSA.

[17] Crémer report (n 13) 3; Gal, M., and Petit, N. (2020). Radical Restorative Remedies for Digital Markets (2021). *Berkeley Technology Law Journal* 37(1), 5-6; OECD, Roundtable on Algorithms and Collusion - Executive Summary (DAF/COMP/M(2017)1/ANN3/FINAL), 26 September 2018, 5; Scott Morton, F., Bouvier, P., Ezrachi, A., Jullien, A., Katz, R., Kimmelman, G., Melamed, D. and J. Morgenstern (2019). Committee for the Study of Digital Platforms, Market Structure and Antitrust Subcommittee, Stigler Center for the Study of the Economy and the State [hereinafter Stigler report], 14.

[18] Crémer report (n 11) 3; Picht, P. G., & Loderer, G. T. (2019). Framing Algorithms: Competition Law and (Other) Regulatory Tools. World Competition, 42(3), 406.

[19] Crémer report (n 11) 2.

[20] When users can multi-home, they might switch to better service providers or use services in parallel, which will increase competition. Interoperability often is a pre-condition for multi-homing and might further allow users to unbundle or establish complementary services. Crémer report (n 11) 23, 83-84. For definitions of multi-homing and interoperability, see footnote 11.

US Congress subcommittee on Antitrust, Commercial Law and Administrative Law addressed the last-mentioned aspect: "Without transparency or effective choice, dominant firms may impose terms of service with weak privacy protections that are designed to restrict consumer choice, creating a race to the bottom".[21] This would further increase the danger of 'tipped markets' in which one company takes the large majority of the market share.[22]

While there is broad consensus that these dynamics can be highly problematic, the exact threshold for a tipped market, or the precise definition of a 'gatekeeper' position arising from such a situation is still up-to-debate.[23] Especially the relation between the legal concept of dominance and the definition of gatekeepers has been subject of a vivid debate among scholars.[24]

Besides these structural challenges, a lot of the competition law debate on digital markets has centered around the issue of *algorithmic* collusion and *tacit* collusion, to be precise. Art 101 TFEU notoriously only condemns collusive arrangements, leaving uncoordinated parallel behavior untouched. For a long time, this was not perceived as a problem since tacit collusion is almost never stable on non-digital markets.[25] However, the increased concentration, transparency, entry barriers and interaction frequency of

---

[21] U.S. House Committee on the Judiciary (2020). Investigation of Competition in Digital Markets. Washington, D.C.: Government Printing Office. The Subcommittee report also mentions manipulative design interfaces, so called dark patterns, nudging consumers into certain choices. Ibid, 53.

[22] Digital Competition Expert Panel (2019). Unlocking digital competition, Report of the Digital Competition Expert Panel, HM Treasury. https://doi.org/10.17639/wjcs-jc14 (accessed 2 March 2021), 3-4 and 88.

[23] Discussions about how to best designate 'gatekeepers' started before the proposal was published and continued in the aftermath of its publication. For a critical view on the gatekeeper definition of Art 3 DMA, see Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T., & von Alstyne, M. (2021). The EU Digital Markets Act - A Report from a Panel of Economic Experts, European Commission, https://ide.mit.edu/wp-content/uploads/2021/02/jrc122910_external_study_report_-_the_eu_digital_markets_act.pdf (accessed 3rd April 2021), 9; Caffarra, C. & Scott Morton, F. (2021). The European Commission Digital Markets Act: A translation, Vox EU CEPR, https://voxeu.org/article/european-commission-digital-markets-act-translation (accessed 3rd April 2021).

[24] For a comprehensive discussion of possible criteria for gatekeepers, see A de Streel, Digital Markets Act – Marking Economic Regulation of Platforms Fit for the Digital Age, Centre on Regulation in Europe Report, 24 November 2020 [hereinafter CERRE DMA report], 35-44. While UK authorities intentionally eschewed the traditional notion of dominance (see e.g., Furman report (n 22) 55), Germany's 10th amendment of its Act against Restraints of Competition (Gesetz gegen Wettbewerbsbeschränkungen, GWB) explicitly includes dominance on one or several markets as part of its definition of 'paramount significance' (§ 19(1) no. 1 GWB), the equivalent to the DMA's 'gatekeepers'.

[25] Beneke, F., & Mackenrodt, M.-O. (2020). Remedies for algorithmic tacit collusion. *Journal of Antitrust Enforcement*, 0(0), 11-12.

online markets make tacit collusion, especially through algorithmic means, a theoretically very plausible scenario.[26] Although the practical relevance of this phenomenon is still disputed,[27] there is no denying that collusion has received considerable attention among scholars and competition authorities.

The list of digital competition buzzwords further includes 'self-preferencing'.[28] Whenever platforms somehow influence consumer choice, e.g., by presenting offers from business customers in a certain order, they have the possibility to favor their own products or services. The *Google Shopping* case established that self-preferential placements are, indeed, not compatible with competition law.[29] Only if the parameters used to rank products are transparent,[30] it will be possible to know whether an online platform is distorting competition by preferencing certain offers, leaving consumers in the dark about the 'trade-offs they are facing', and hence inhibiting competition in a significant manner.

Existing laws often fall short in addressing these issues, both in the EU and the US. For instance, the US Subcommittee on Antitrust has noted that 'some of these business practices are a detriment to fair competition, but they do not easily fit the existing categories identified by the Sherman Act, namely 'monopolization' or 'restraint of trade'. The DETOUR bill, adopted on a bi-partisan basis by the US Congress in 2020, can thus be read a means to 'prohibit the usage of exploitative and deceptive practices by large online operators and to promote consumer welfare in the use of behavioral research

[26] Bundeskartellamt, & Autorité de la Concurrence. (2019). Algorithms and Competition, https://www.bundeskartellamt.de/SharedDocs/Publikation/EN/Berichte/Algorithms_and_Competition_Working-Paper.pdf?__blob=publication File&v=5 (accessed 20th October 2020), II.

[27] For a critical view, see Mehra, S.K. (2017). Robo-Seller Prosecutions and Antitrust's Error-Cost Framework. *CPI Antitrust Chronicle*, https://www.competitionpolicyinternational.com/robo-seller-prosecutions-and-antitrusts-error-cost-framework/ (accessed 3rd April 2021).

[28] For an in-depth discussion, see Geradin, D., & Katsifis, D. (2020). "Trust me, I'm fair": Analysing Google's latest practices in ad tech from the perspective of EU competition law. *European Competition Journal*, 16(1), 11–54. Self-preferencing was at the heart of the Microsoft saga (on which, see Jennings, J. P. (2006). Comparing the US and EU Microsoft Antitrust Prosecutions: How Level Is the Playing Field. *Erasmus Law and Economics Review*, 2, 71–86.)

[29] In the Google Shopping case, self-preferential placements were deemed not compatible with competition law: Google Search (Shopping) Case C(2017) 4444, 27 June 2017, paras 9, 10 of summary decision. See Ackman, P. The Theory of Abuse in Google Search: A Positive and Normative Assessment Under EU Competition Law, in *Journal of Law, Technology & Policy*, (2) 301-372.

[30] Picht, P. G., & Loderer, G. T. (2019). Framing Algorithms: Competition Law and (Other) Regulatory Tools. World Competition, 42(3), 416.

by such providers'.[31]

In the debate surrounding the DSA and DMA proposals, general shortcomings of EU transparency and competition rules when dealing with opaque online platforms have been highlighted.[32] Relying exclusively on Arts 101, 102 TFEU might mean that the Commission could only act on a case-by-case basis, for a very limited set of platforms, ex post, and only after lengthy investigations.[33]

Given that Section 2 of the Sherman Act (15 U.S.C. § 2) and Section 7 of the Clayton Act (15 U.S.C. § 18) are even more narrow than comparable EU competition norms, their suitability to achieve a satisfactory level of fairness and competition effectively create such a situation might be even more limited, especially regarding effective remedies.[34]

In light of these interconnected challenges for consumer protection and competition, a consensus has been reached on the need new ex ante rules to

[31] See Deceptive Experiences To Online Users Reduction Act' or the 'DETOUR Act', H.R. 8975, 116th Congr., (2020). https://www.govinfo.gov/content/pkg/BILLS-116hr8975ih/pdf/BILLS-116hr8975ih.pdf The Bill has not been resumed by the newly established administration (2021).

[32] The Crémer report points out several criticalities: (1) not all gatekeepers enjoy a dominant position in the sense of Art. 102 TFEU; (2) the relevant market might be substantially harder to define than in non-digital cases; (3) not every problematic practice has a demonstrable effect on the relevant market. The authors conclude that greater emphasis should be put on the theory of harm, instead. Crémer report (n 13) 3-4. Moreover, digital markets are often moving at a rapid pace, which is not necessarily a characteristic they share with competition law. Hence, there are concerns whether competition law could be applied with the necessary speed to address urgent competition needs. CERRE DMA report (n 24), 59; Recital 5 DMA.

[33] This is not to say that existing competition norms are generally useless to address novel competition issues; in fact, EU case law has shown the opposite. E.g., Google Search (Shopping) Case C(2017) 4444, 27 June 2017; Google Android Case C(2018) 4761, 18 July 2018. On the national level the German competition authority has taken action against certain data collection practices of Facebook (see Bundeskartellamt, Decision B6-22/16, 6 February 2019). Also with a focus on data collection, the Commission has opened an investigation against Amazon in 2019 (European Commission, Antitrust: Commission opens investigation into possible anti-competitive conduct of Amazon, 17 July 2019, https://ec.europa.eu/commission/presscorner/detail/en/ip_19_4291, accessed 16 February 2021). For older cases, see Waller, S.W. (2012). Access and Information Remedies in High-Tech Antitrust, Journal of Competition Law and Economy 8(3), 576. Regarding the US, some argue in a similar vein: "This is not to say that the use of the antitrust laws should be abandoned. If history is a guide, there is a meaningful possibility that antitrust enforcement activities will produce value commensurate with their costs.". Wheeler, T., Verveer, P., & Kimmelman, G. (2020). New Digital Realities: New Oversight Solutions in the U.S. Harvard Kennedy School Shorenstein Center, August 2020, https://shorensteincenter.org/wp-content/uploads/2020/08/New-Digital-Realities_August-2020.pdf (accessed 5 February 2020), 26.

[34] Wheeler et al. (n 33) 24-26.

complement the existing legal framework.[35]

B. How the DMA and DSA Proposals Respond to the Quest for New Pro-Competitive Rules

The European Commission's vision of these rules was first outlined in three separate inception impact assessments,[36] that were open to consultation by stakeholders. As a result, the DMA and DSA normative proposals were adopted that encapsulate such discussion. In the following, we will be briefly present their content.

Both proposals address negatives externalities and asymmetric information. However, the DSA is more consumer protection-focused, while the DMA's goal is to '*ensur[e] fair and open digital markets*'[37] by applying asymmetric rules against large online platforms designated as '*gatekeepers*', which are addressed with a list of does and don'ts.[38]

The DSA applies to all 'intermediary services' (Art 1), while the scope of the latter is limited to 'core platform services' offered by 'gatekeepers' as defined in Art 3 DMA. It replicates the GDPR's approach to applicability, hence applying to all services provided to EU citizens 'irrespective of the place of establishment of the providers of those services'[39]. This is a characteristic it shares with the DMA[40].

However, the DMA is much more limited in scope since its goal is not to lay down a fundamental framework, but rather to complement existing competition law norms with respect to a very specific set of market players (the gatekeepers).[41] Regarding its *ratione materiae*, the DMA has a more

[35] See for instance Algorithm Watch (2020), Governing Platforms – Final Recommendations, available at https://algorithmwatch.org/wp-content/uploads/2020/10/Governing-Platforms_DSA-Recommendations.pdf (accessed 17 February 2021), 1; SB Micova & A de Streel, Digital Services Act – Deepening the Internal Market and Clarifying Responsibilities for Digital Services, Centre on Regulation in Europe Report, 2 December 2020, https://cerre.eu/publications/digital-services-act-responsibility-platforms/ (accessed 16 February 2021) [hereinafter CERRE DSA Report], 39; European Parliament, Draft Report with recommendations to the Commission on Digital Services Act: Improving the functioning of the Single Market (2020/2018(INL)), Committee on the Internal Market and Consumer Protection, 24 April 2020, 5(no 15).

[36] See above (n 15).

[37] Recital 5, DMA.

[38] In fact, the DSA amends the e-commerce directive (2000/31/EC), the DMA seeks to complement EU competition rules (mostly Art 101, 102 TFEU).

[39] Art 3 DSA.

[40] Art 1(2) DMA.

[41] See above (n 7).

limited scope (the core platform services listed in Art 2(2));[42] to be in line with its much more specific objective of 'ensuring contestable and fair markets … *where gatekeepers are present*'[43]. Given the centrality of the 'gatekeeper' concept, different understandings of this term could cast doubts at whether all stakeholders mean the same when they express their support for additional rules for 'gatekeepers'.

Turning to some of the substantive rules for said gatekeepers, they will have to grant data access to business users for the data they generated on the platform (Art 5(h) DMA) and allow for multi-homing (Art 5(e) DMA) and interoperability (Art 5(f) DMA). Self-preferencing will be explicitly prohibited (Art 6(1)d DMA), as well as preventing users from un-installing certain, pre-installed tools (Art 5(b) DMA). With a view to advertisement markets, Art 5(g) DMA would oblige gatekeepers, with respect to their core platform services within the meaning of Art 3(7) DMA, to provide information about pricing and performance measuring tools allowing them to assess how satisfied they are with the advertisement product they are paying for (Art 6(1)g DMA).[44]

With a view to the connection between transparency and competition mentioned above, a look at the DSA is helpful to complete the picture. Its Arts 12(1), 13, 23, 24, 29, 33 establish comprehensive, but differentiated disclosure and reporting duties regarding ranking, advertisement, and content moderation practices.[45] The more pronounced transparency obligations for very large online platforms within the meaning of Art 24 DSA reflect the differentiated approach the Commission took for the design of the DSA, explicitly mentioned in Recital 39 of the proposal.

To sum up, this section has shown that certain characteristics of digital services might impair competition. It has also sketched out the debate on why existing rules might be insufficient to prevent this and how the DMA and DSA proposals of the European Commission seek to change this.

II. A Computational Analysis of the DSA and DMA Consultation Process

[42] Note that electronic communication network and services markets are exempted from the scope of the proposal (Art 1(3) DMA).

[43] Art 1(1) DMA, emphasis added.

[44] While these obligations are rather specific, Art 10 DMA would open the door to add further duties in the future if a market investigation pursuant to Art 17 DMA identified a need to do so for the sake of safeguard fair competition.

[45] Just to name those for very large online platforms within the meaning of Art 25 DSA, Art 29 DSA would entail the obligation to provide information on the use of recommender systems and their parameters. Art 33 sets out comprehensive transparency obligations for very large online platforms.

In motivating its twin proposals, the Commission reports that the 'vast majority of respondents' in the consultation process '*considered that dedicated rules on platforms should include prohibitions and obligations for gatekeeper platforms*'.[46] More specifically, in the DMA this 'majority' believed that the '*the proposed list of problematic practices, or "blacklist", should be targeted to clearly unfair and harmful practices of gatekeeper platforms*'.[47] In the DSA, the quest for *'algorithmic accountability and transparency audits, especially with regard to how information is prioritized and targeted*' online comes from '*a wide category of stakeholders*', and is particularly voiced by 'civil society and academics'.[48]

In this chapter, we ask whether ex ante rules for very large platforms and gatekeepers are what stakeholders demanded for in the consultation process, and whether their actual wording in the DSA and DMA proposals reflects the way each group use the relevant terms. This is a relevant step, as it is important that the addressees of such duties (typically digital firms) and the beneficiaries (individuals, micro and small organizations using platforms) agree on their meaning.

To do so, we use computational text analysis techniques to analyze the replies to the questionnaires and position papers submitted by stakeholders to the EU consultation process for both proposals.

A. Data and Methodology

To analyze the said documents, we created a special scraper algorithm, which allowed us to download all the files automatically, convert them into text, and split them into three clusters. In doing this, we relied on the Commission's categorization, based on the organization size of the feedback contributors.[49] We then aggregated the different sub-categories into three corpuses, based on typology and dimension of the feedback contributor:

Corpus A (individuals and micro organizations),

---

[46] DMA, above (n 2) at 8 (summarizing the results of stakeholder consultations and impact assessments).

[47] Ibid.

[48] DSA, above (n 15) at 9. See also Algorithm Watch (n 35) 1; CERRE DSA report (n 35) 39; European Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust (COM(2020) 65), 19 February 2020, 15.

[49] The Commission distinguishes between (1) individuals, micro (< 10 employees), (2) small (< 50 employees), (3) medium (< 250 employees), and (4) large (250 or more) organizations as well as between different types of feedback contributors (i.e., in the DSA, p. 8 you find that respondents are: the general public (66%), companies/businesses organizations (7.4%), business associations (6%), and NGOs (5.6%) authorities (2.2%), academic/research institutions (1.2%), trade unions (0.9%), and consumer/environmental organizations (0.4%)). DSA, above (n 2) 8.

Corpus B (small companies/organizations), and
Corpus C (medium and large companies/organizations).

This clustering scheme was informed both by a preliminary analysis of our data, which is explained in more detail in the Appendix (Annex 1). Interestingly and importantly, it showed that the classical 'small and medium enterprises' group was not equally applicable to our data, since the feedback contributors do not only comprise enterprises, but a more diverse set of actors.

To discern differences in the use of certain key terms across stakeholder groups (i.e., a different semantic understanding of identical terms), we leveraged Word Embedding Models to quantify evidence of such differing understandings. This technique has already been used in various Natural Language Processing tasks,[50] and recently also in the Computational Law literature. More specifically, it has been demonstrated to be very powerful and useful in providing insights into latent differences of how language is used in legal scholarship[51] and commences with some applications in computational antitrust literature.[52]

The core of this technique consists in training a special neural network to convert each word contained in a corpus of texts into a vector, i.e., a set of numbers.[53] While a simple algorithm would require to formulate explicit rules to somehow approximate the semantic meanings of words, ML (or the neural network, to be precise) learns the implicit rules directly from the data we feed it. The resulting vectors are based on the frequency words occur next to each other, meaning their relative positions in each phrases of the corpus

[50] For instance, word embeddings have been used by sociologists to investigate the meaning of the term 'class', to predict conflict, or to classify documents. Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. American Sociological Review, 84(5), 905–949; Kutuzov, A., Velldal, E., & Øvrelid, L. (2017). Tracing armed conflicts with diachronic word embedding models. Proceedings of the Events and Stories in the News Workshop, 31–36; Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in Twitter election classification. Information Retrieval Journal, 21(2), 183–207.

[51] Nyarko, J., Sanga, S. (2020). A Statistical Test for Legal Interpretation: Theory and Applications, 25 November 2020, https://juliannyarko.com/wp-content/uploads/other/nyarko_sanga_legal_interpretation.pdf.

[52] For a discussion of computational techniques in antitrust law and their implications, see the following: Schrepel, T. (2021). Computational Antitrust: An Introduction and Research Agenda. Stanford Computational Antitrust, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3766960 (accessed 12 May 2021); Massarotto, G., & Ittoo, A. (2021). Gleaning Insight from Antitrust Cases Using Machine Learning. Stanford Journal of Computational Antitrust, 1, 16–37.

[53] The Neural Network in particular is a LSTM (Long-Short Term Memory Network). See Hochreiter, S., Schmidhuber, J. (1997). Long Short-term Memory. Neural Computation 9(8):1735-80.

and the correlation between words. The stronger two words are correlated (in their occurrence – and so in their semantic meaning) [54] in the corpus the model was trained in, the closer the corresponding vectors will be located to each other.

However, models trained on different corpuses are not directly comparable as they depend on the corpus the model was trained on. The vector of a single word alone does not provide meaningful insights. To test if there is evidence of a different semantic use of the same words between two texts, we assess the distance between vectors from two different corpuses corresponding to the same words. To align them, we transform the two models geometrically.[55] This allows to understand how a vector in one corpus relates to the vector of another corpus. After the transformation, the vectors of the two aligned corpuses are comparable. Therefore, for each corpus we trained a different word embedded space, and we aligned each pair of words occurring in both corpuses through the means of Unsupervised Vector Space Alignment.[56]

Having aligned our three corpuses, we are able to compute the distance between the same terms from different corpuses. However, how do we know that the distances we find are not just random, but actually based on substantial uses and understandings between stakeholders? To see if there is evidence for a statistically significant semantic difference between the use of a term between the different stakeholder groups, we must perform a statistical test (see Appendix, Annex 2 for detailed description). Otherwise, we would not be able to tell whether the differences we find between corpuses are

[54] This is based on the 'distributional hypothesis', which assumes that words which frequently occur together are usually also semantically related. While this approach might seem too simple to capture complex semantic meanings, the success of algorithms relying on it suggests that the claim has some merit. Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2017). Comparative study of LSA vs Word2vec embeddings in small corpora: A case study in dreams database. Consciousness and Cognition 56, 178–187.

[55] To perform this transformation, we used a 'control vocabulary', containing a list of words that we can safely assume that share the same semantical meaning. The list of 1,189 words we used is, in fact, composed mainly of numbers and stop-words (like e.g. 'the'). We are thankful to Professor Julian Nyarko from Stanford University for providing us with a first list of control keywords, to which we further added almost 2.000 stop-words and numerals we took from the different corpuses. We manually selected our control vocabulary. We used the glossaries contained in all EU directives and regulations recalled in the DSA and DMA proposals and published in the OJUE. Further, we manually coded the questionnaires (used in the consultation) and selected terms of interest.

[56] We used a special algorithm provided by Facebook in the library FastText. (https://github.com/facebookresearch/fastText), used in Python. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information, http://arxiv.org/abs/1607.04606 (accessed 22 January 2021).

actually relevant or merely signs of random, non-semantic differences between our corpuses. By modeling the theoretical distribution of non-semantic differences for the terms in our corpus, we can compare the distance we would expect to see if there was no semantic difference with the distance we observe. In this way, we can conclude with a certain level of confidence that the distance we observe between our corpuses is more than just random or syntactical.

In order to gain deeper insights into possible reasons for a semantic difference in the use of key words between different corpuses, we also leveraged the tool of Sentiment Analysis,[57] applied to sentences of the two compared corpuses where the specific key word appears. Sentiment analysis is a Natural Language Processing technique, which classifies a sentence, or a paragraph based on the use of specific words and their location inside the text, giving as a score a value of the positive or negative sentiment inside the particular text. These two values are aggregated to a compound value, which gives a score to the overall sentiment ranging from -1 (totally negative) to +1 (totally positive).

B. Results and Discussion: Different Groups, Different Uses?

With these tools at hand, we were able to find a significant difference for 1,865 word pairs between Corpuses A and C. Between Corpuses A and B we found 2,184 statistically significant differences and 1,113 between stakeholder groups B and C.[58]

In the following, we will only discuss the most interesting differences we found, that are relevant to (1) key actors and structure of digital markets, (2) their anticompetitive conduct, and (3) the identified remedies and ex ante rules. The results are summarized in Tables 1 to 3 respectively.

Summary of results

Table 1: ***Key actors and market structures***

Term	Distance AB	Distance BC	Distance AC	Close words A	Close words B	Close words C

[57] Khan, F. (2018). Sentiment Analysis of Twitter Data. *International Journal of Engineering Research 6*(15).

[58] Note that many of these words are not of particular interest for us as they might be specific to a position paper of a certain company (e.g. 'Gmail' in Google's submissions). However, some of the key buzzwords surrounding competition law and new ex ante remedies show statistically significant differences.

Dominant	0.515 (0.65)	0.797 (0.50)	1.530 (0.02)**	Self-preferencing		Policymakers, quasi-judicial
Gatekeepers	1.064 (0.22)	1.367 (0.11)	1.486 (0.03)**	Content		Unsatisfactory
Monopolistic	1.023 (0.25)	1.473 (0.06)*	1.631 (0.00)**	Discouraging, profitability	Vulnerability, linking	Higher-cost, welfare
Monopolization*	1.323 (0.109)	1.432 (0.078)	1.609 (0.00)**	Endangers, non-dominant, data		Operations
Newcomers	1.469 (0.03)**	1.571 (0.04)**	1.192 (0.18)	Non-existent, none	Tech	Start-ups, destroyed

**'Monopolization' was used a relatively small number of times (only 285), compared to other reported words. We decided to include it to show the semantic distance with neighboring concepts.*

Table 2: ***Anticompetitive conduct***

Term	**Distance AB**	**Distance BC**	**Distance AC**	**Close words A**	**Close words B**	**Close words C**
Abusive	1.216 (0.14)	1.162 (0.20)	1.543 (0.04)**	Competing		Tacit, monopoly
Collusion	1.590 (0.01)***	1.420 (0.08)*	1.396 (0.08)*	policing	tacit, data	
Coordinated	1.395 (0.06)*	1.110 (0.23)	1.541 (0.02)**	Cooperation, dialogue	Tacit, legitimate	Future-proof, EU-level
Pro-competitive	1.445 (0.04)**	1.295 (0.13)	1.022 (0.29)	comparison	user-friendly	
Ranking	1.182 (0.15)	1.644 (0.02)**	1.452 (0.05)**		guidelines, improve, oversight	appearance, disclosing
Self-favoring	.	.	1.589 (0.00)**	combating		debates
Self-preferencing	1.393 (0.04)**	1.457 (0.07)*	1.190 (0.18)	monopolizing	over-regulation, reports	
Tacit	1.439 (0.04)**	0.923 (0.40)	1.274 (0.13)		threatens	
Tipping	1.509 (0.03)**	1.584 (0.03)**	1.213 (0.17)		bottleneck, nudge	unwanted

Uncompetitive	1.478 (0.03)**	1.199 (0.18)	1.499 (0.03)**	single-sided, state-run	EU-commission, institutions	

Table 3: ***Remedies and ex ante antitrust reform***

Term	**Distance AB**	**Distance BC**	**Distance AC**	**Close words A**	**Close words B**	**Close words C**
Blacklist	1.402 (0.05)	1.063 (0.27)	1.580 (0.01)**	functionalization		Dominance-based, problem
Data-sharing	1.566 (0.02)**	.	.	recycled	differentiation	
Disproportionally	1.590 (0.01)**	1.044 (0.28)	1.154 (0.20)	Public-interest, bans	Ensure, competition, incentives	
Interoperability	1.665 (0.01)**	1.150 (0.21)	0.756 (0.49)	reliability, trustworthy	licensing	
Overregulated	1.500 (0.03)**	.	.	Sellers	Tax-like	
Pro-competitive	1.445 (0.04)**	1.295 (0.13)	1.022 (0.29)	comparison	user-friendly	
Underenforcement	1.463 (0.03)**	1.283 (0.13)	1.513 (0.02)**	Complaints, unbureaucratic	consensus	Misconceptions, improvements
Unregulated	1.361 (0.07)*	1.566 (0.04)**	1.822 (0.00)***	not-sufficient		mitigation
Welfare	0.989 (0.26)	1.543 (0.04)**	0.626 (0.65)		Economic, rights	mobility

Note: The 'Distance' columns report the distance between the vectors of the same words for each corpus pair with the respective p-value in parentheses. A grey field indicates that a word was not used in both of the respective corpuses. The 'Close Words' columns shine a light at some of the concepts that were closely related with the term in question in the corpuses for which there was a statistically significant distance between the terms. The asterisks indicate significance at a 0.001 (***), 0.05 (**), and 0.1 (*) level, respectively.

1. Key Actors

Figure 1: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpuses

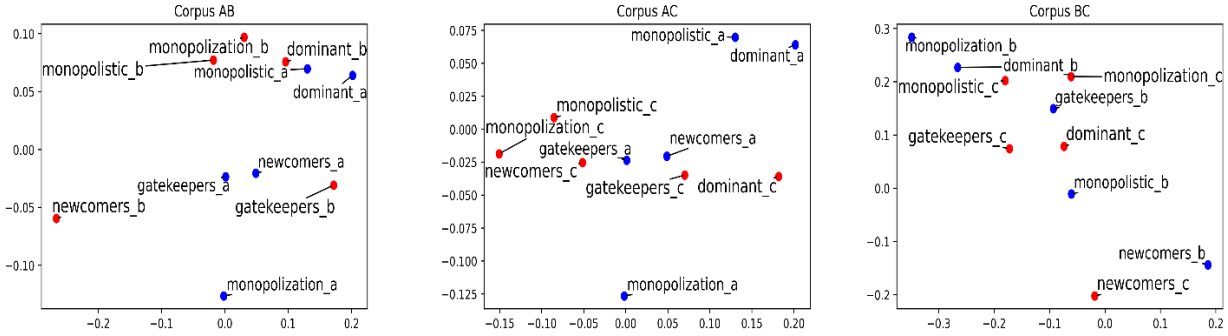
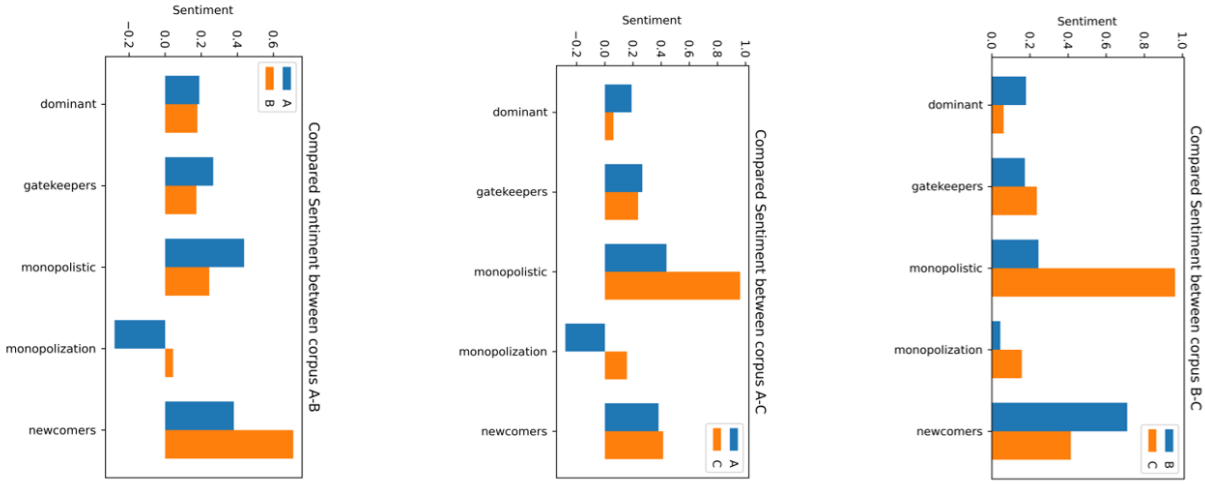


Figure 2: Sentiment Analysis. Compared Corpuses



Starting with terms related to key actors, and especially the applicability of competition law and the DMA, we see a statistically significant difference between micro companies/organizations and the medium/big companies and organizations, for an absolutely central term: 'gatekeepers'. This can be linked back to the Commission's summary of the consultation, where it pointed out that some stakeholders considered the concept of a gatekeeper too broad.

Here, our computational analysis confirms the qualitative assessment of the Commission. The question of how 'gatekeeper' should be defined was hotly debated[59] and it seems like the differing standpoints also translated into a different use of the term 'gatekeeper' in the position papers for all three corpuses.

Another term that has risen in importance in the recent debates on digital markets is 'newcomers', the counterpart of 'gatekeepers'. We find that the

[59] See DSA, above (n 35), 8.

term is used differently between both Corpus A and B as well as between Corpus B and C. Interestingly, it is closely related with 'non-existent' in Corpus A and 'destroyed' in Corpus C, which seems to hint at the difficult standing of small tech start-ups on certain digital markets.

Remarkably, we also find a significant difference for the term 'dominant', with Corpus A showing a close connection to the term 'self-preferencing'. This is noteworthy with a view to Art 102 TFEU, which applies only to firms holding a 'dominant' position. This finding suggests that that not only new competition concepts, but also established ones are perceived differently, in this case by micro-organizations/individuals and medium/big organizations.

Unlike 'dominant', EU competition law does not make use of the term 'monopoly' or related concepts. However, they are frequently used in debates on competition law, including those on the DSA and DMA. Our analysis shows that there is no perfect agreement between different stakeholder groups concerning the precise content of terms like 'monopolistic' or 'monopolization'. While the novelty of a term like 'gatekeeper' might to some extent explain differences in its use, it is surprising that more established terms are used in an equally non-uniform way. This could be due to the fact that many respondents are non-Europeans and therefore not used to the EU competition jargon. However, as the close words analysis for 'monopolization' reveals, this finding might also be related to the debates on how even non-dominant firms might gain a position that enables them to, e.g., effectively bar start-ups ('newcomers') from entering the market. The fact that also 'data' is closely related to 'monopolization' could be interpreted as a hint of how new technologies can be a drive for economic giantism, while introducing uncertainty around relatively well-established terms.

Nevertheless, it needs to be noted that we did not find a significant difference for terms like 'platform', 'market power' or 'dominance', which should be encouraging.

2. Anti-Competitive Conduct

Figure 3: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpuses

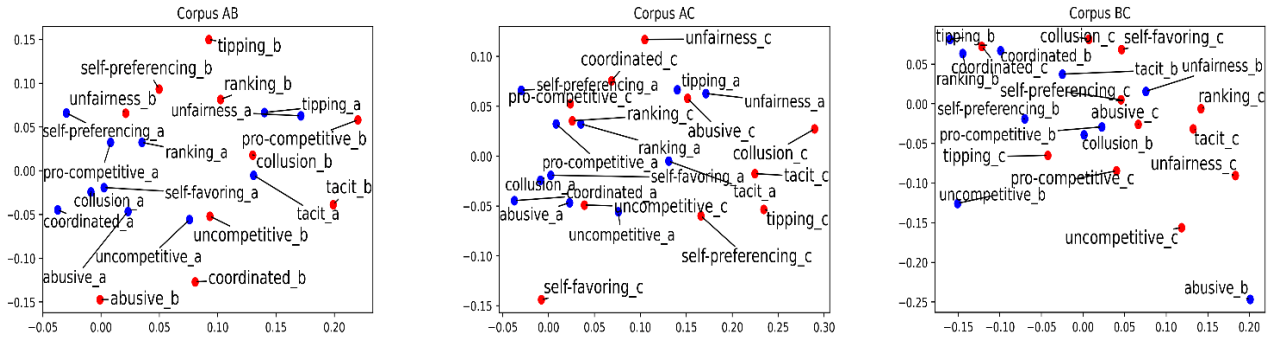
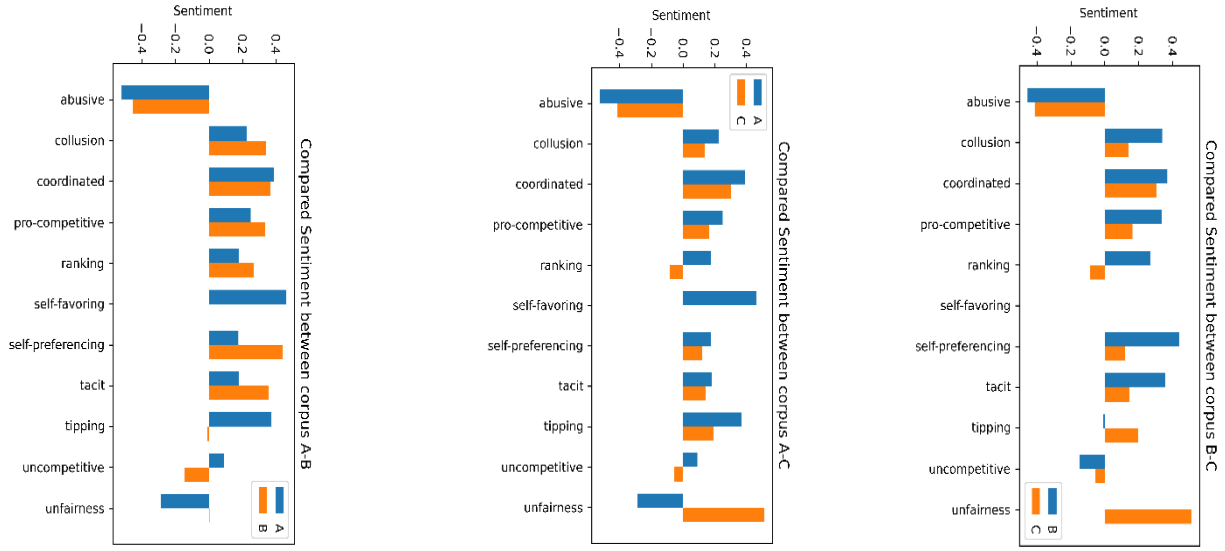


Figure 4: Sentiment Analysis. Compared Corpuses



Moving on to terms related to potentially *Anti-Competitive Conduct*, it is remarkable to see that general terms like 'uncompetitive' and 'pro-competitive' are not used in the same way. The same holds true for the more drastic, but also commonly used term 'abusive'. Interestingly, 'pro-competitive' is closely associated with 'user-friendly' for small companies and organizations (B), which could be interpreted as a hint that this group of stakeholders, in particular, perceive a strong link between consumer protection and friendliness (or fairness, in the DMA and DSA jargon) and competition. Individuals and micro-organizations (A), on the other hand, connect 'pro-competitive' with the term 'comparison. This could emanate

from the centrality of comparing offers to boast competition and innovation, and seems to confirm the importance of provisions like Art 6(1)d DMA[60] or Art 29 DSA.[61]

Strictly related to that, is the statistically significant distance we find between the vectors of 'self-preferencing' for corpuses A and B as well as for B and C. For corpus pair AC, the alternative term 'self-favoring' is used differently as well, suggesting that it is the general concept behind these two terms that seems to be still elusive. In corpus A, 'self-preferencing' is most closely related to 'monopolizing', indicating that the anti-competitive outcomes of self-preferencing are a key concern for individuals and micro-organizations.

Small companies, on the other hand, associate this practice with '*over-regulation*' and '*reports*'. This is interesting not only with a view to the prohibition on self-preferencing in Art 6(1)d DMA, but also in light of the comprehensive reporting duties on rankings in the DSA (Arts 12, 29 DSA). Our results could be interpreted as a clue that small companies and organizations might fear comprehensive transparency and reporting duties, thus highlighting the need for a differentiated approach.[62]

Another anti-competitive practice which yields a significant distance between corpuses A and B is 'collusion'. For small businesses and organizations (B), this keyword is closely related to '*tacit*' which well-reflects the many debates on 'tacit collusion', driven both by the doctrine[63] as well as competition authorities,[64] and lawmakers.[65] Note that the term 'tacit' itself is used in a similarly idiosyncratic manner, which might reflect the debate on what exactly constitutes an agreement or otherwise unlawfully coordinated behavior. The term 'coordinated' is also used differently by different stakeholders, although only small companies seem to connect it with tacit collusion. Interestingly, another term that shows up in the vicinity of 'collusion' is '*data*', which is in line with the conclusion drawn by many

[60] Art. 6(1)d DMA regards self-preferencing in the context of ranking services (a practice addressed in the *Google Shopping* case, above (n 29).

[61] Above (n 45).

[62] However, such a proportional approach is not easy to find, especially if there is no consensus on what 'disproportionally' burdensome provisions look like. This seems to be the case at least for corpus pair AB.

[63] See e.g. Ezrachi, A. & Maurice E.S. (2020) Sustainable and Unchallenged Algorithmic Tacit Collusion, *Northwestern Journal of Technology and Intellectual Property* 17, 217-59.

[64] See e.g. Bundeskartellamt & Autorité de la Concurrence, Algorithms and Competition, November 2019, https://www.bundeskartellamt.de/SharedDocs/Publikation/EN/Berichte/Algorithms_and_Competition_Working-Paper.pdf?__blob=publication File&v=5 (accessed 25 February 2021).

[65] Crémer report (n 13) 68.

experts regarding the central role of data availability as an enabling factor of tacit collusion.[66]

It is also worth noting that no reference is made in the DMA to 'data' or 'algorithmic collusion', even though both locutions were referred to in the consultations. Given that other close relations exist with the words 'harms' and 'barred', we could deduce a rather negative attitude towards the emergence of collusion in corpus B. Generally, collusion is relevant with a view to existing competition law (esp. Art 101 TFEU), but was not picked up in the DMA, despite the long debate.

A potential avenue for anti-competitive conduct of online platforms that is being addressed by both the DMA (Art 29 DSA) and the DMA (Art 6(d) DMA) are '*rankings*', for which we find a statistically significant distance between corpuses A and C. A similar discrepancy was found for '*tipping*', which is an important concept related to the need for *ex ante* rules[67]. In the small companies/organizations corpus, 'tipping' is surrounded by the terms 'bottleneck' and 'nudge'. This mirrors the debate around gatekeepers, which are often considered to gain their dominant or quasi-dominant role from their position as bottlenecks, nudging consumers into certain choices.[68] Interestingly, our sentiment analysis (Fig. 4) shows that 'tipping' is viewed more favorably by medium and big companies/organizations (0.338) than by small companies/organizations (-0.296). The negative attitude of small companies and organizations towards 'tipping' might reflect the potentially even higher costs of uncompetitive markets small businesses face, which might often depend on uncontestable platform markets.

[66] See, for instance, Crémer report (n 11) 8.

[67] See footnote 22.

[68] There has even been empirical evidence for problematic nudges, so called 'dark patterns'; Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW)*, 81:1–81:32. Concerning bottlenecks, the UK Furman report incorporated the concept in its 'significant market status' threshold, which is comparable to the 'gatekeeper' designation in the DMA. Digital Competition Expert Panel (n 22) 10.

3. Regulatory Reform and Remedies

Figure 5: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpuses

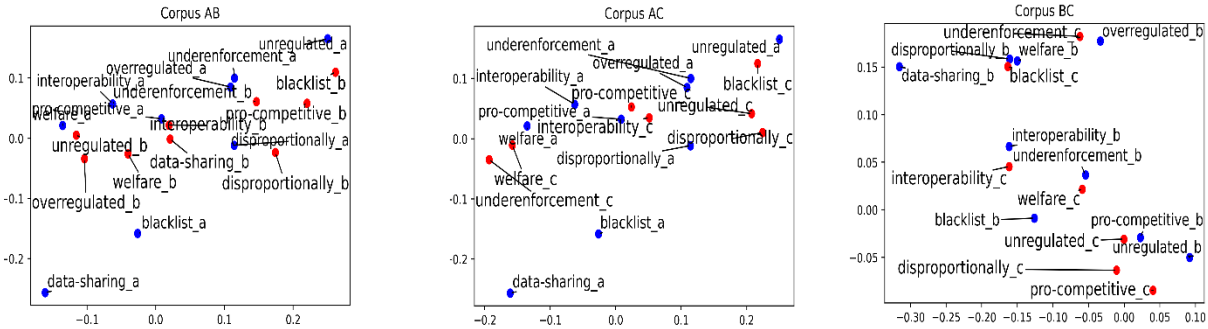
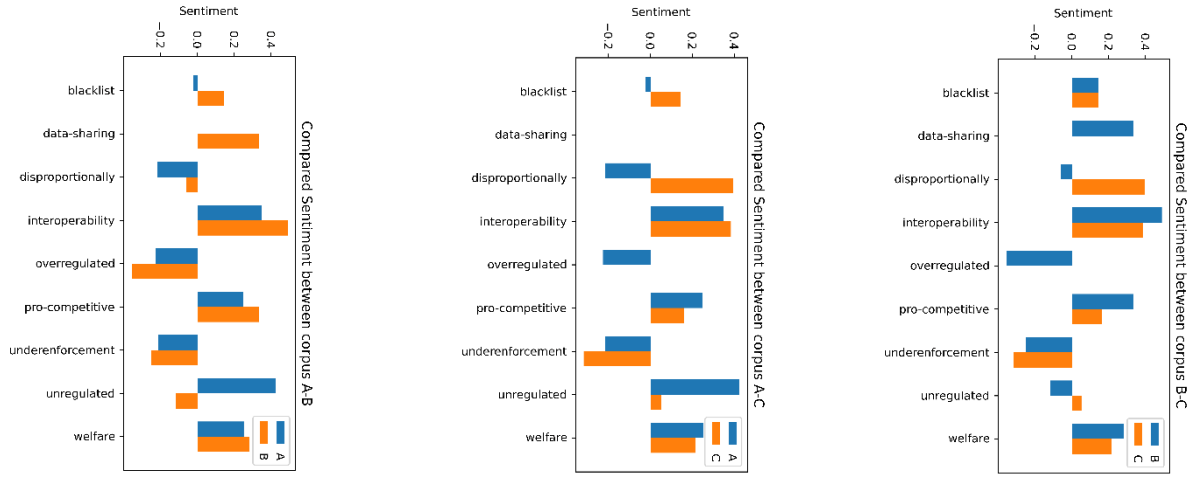


Figure 6: Sentiment Analysis. Compared Corpuses



Having discussed concepts related to competition challenges on digital markets, the logical next question seems to be: What to do about it? Regarding the general level of regulatory intensity, there seems to be no perfect alignment on the meaning of 'overregulated', with small organizations appearing to fear 'tax-like' measures. At the same time, at least micro-organizations and individuals look to perceive a lack of regulation as 'not sufficient'. Generally, for the term 'unregulated' we find significant differences between all corpuses. In a similar vein, we find a considerable distance between the vectors of 'underenforcement' for Corpuses A and B as well as A and C. While micro-organizations and individuals seem to emphasize the importance of 'unbureaucratic' procedures, medium and big organizations/companies interestingly speak of 'misconceptions' and 'improvements'. While the precise meaning of these associations remains

obscure, it is interesting to note that not only individuals and micro-entities but also bigger organizations seem to demand some kind of reform.

However, regarding the concrete measures of reform proposed by the Commission, we find that stakeholders have different understandings of three key concepts. First, the term 'blacklist', which is highly relevant for the DMA,[69] is not used in the same way by micro entities as by medium/big organizations. Interestingly, the former mentioned the concept far less than the latter (1,972 occurrences in Corpus A, compared to 36,161 in Corpus C). This disparity could be explained by a certain sense of alarm on the side of larger entities that is hinted by the associated term 'problem'. While this interpretation is highly tentative and would need to be confirmed by further analysis, the second closely associated term, 'dominance-based' is a bit more telling and highly noteworthy. In fact, it might reflect the demand to set the threshold of application for new, stricter rules at the 'dominance' threshold.[70] Linking this finding with our results for the terms 'gatekeeper' (the approach chosen by the Commission) and 'dominant', it seems reasonable to conclude that the disagreement on the applicability *ratione personae* of new rules is also reflected in the use of certain pivotal terms related thereto. Furthermore, our sentiment analysis (Fig. 6) shows that 'blacklists' are seen more favorably by small, medium, and large companies than by individuals/micro entities. This might explain some of significant differences we find in the use of the term.

Second, moving from the 'don'ts' of the blacklist part of the DMA to the positive obligations complementing these prohibitions, we find two noteworthy differences between Corpuses A and B. For instance, 'interoperability' was used in a substantially different way, which might be a reflection of the debate on how far-reaching interoperability obligations should be.[71] Interestingly, the focus on 'reliability' and trustworthiness in the closely associated words in Corpus A matches the spirit of Art 6(1)(c) DMA, which grants gatekeepers the opportunity to restrict interoperability to the extent necessary to ensure 'the integrity of the hardware or operating system provided'.

Lastly, we found significant differences in the use of the term '*data-sharing*'. While individuals and micro-organizations (A) emphasize the possibility to 'recycle' the data gatekeepers are obligated to share under, e.g.,

[69] To be specific, Art 5(1)(a), (d)-(f) DMA and Art 6(1)(a), (d)-(e) DMA 'blacklist' certain actions.

[70] See e.g., CERRE, DMA report (n 24), 52.

[71] Currently, Art 6(f) DMA only covers ancillary services, which many would like to see changed. Stolton, S. (2021). EU SMEs in bid for greater interoperability in Digital Markets Act. *Euractiv*, https://www.euractiv.com/section/digital/news/eu-smes-in-bid-for-greater-interoperability-in-digital-markets-act/ (accessed 13/05/2021).

Art 5(1)(g) and Art 6(1)(g) DMA, small companies and organizations (B) heed the need for 'differentiation'. Potentially, this association is an expression of a fear of being overburdened with costly, potentially disproportional data-sharing duties. However, what exactly 'disproportionally' burdensome remedies look like seems to be a matter of debate for itself, given that the term shows a significant difference between corpuses A and B.

C. What Lessons Can We Draw?

Using computational tools to analyze the DSA and DMA consultation feedback documents allowed us to reach three results. First, we found that considering small and medium entities as a homogenous group of stakeholders might not always be recommendable, at least not for natural language data. In our computational analysis, it turned out that medium organizations are more comparable to big contributors than small ones. This highlights the need for data-driven procedures when determining the optimal units of analysis. As outlined below, there are several computational tools the Commission could use to respond to this need.

Second, our algorithmic analysis has shown that there are statistically significant differences between stakeholder groups in the use and understanding for some key concepts of competition policy. On the one hand, our analysis reproduced some of the results the Commission outlined in the summaries of the consultations. For instance, the differences in the terms 'gatekeepers' or certain remedial strategies between different groups can be linked back to the debate on these concepts in the sense that they were not entirely uncontroversial.[72] On the other, we spotted substantial differences between stakeholders for terms the Commission perceived as uncontroversial (for example, the Commission concluded that '[t]he large majority of stakeholders believed that the proposed list of problematic practices, or "blacklist", should be targeted to clearly unfair and harmful practices of gatekeeper platforms'; while we found that the use and understanding of 'blacklist' differs significantly among different feedback contributor groups). The same holds for some important anticompetitive practices (such as 'collusion', 'tipping', or 'self-preferencing'). It follows that the consensus the Commission identified over the proposed ex ante measures might not be as unanimous as it seems at first sight: although stakeholders might say the same, they could mean different things.

The tools we used in this paper allow us to discern such differences. Hence, when analyzing stakeholder submissions, an examination like the one

[72] DMA proposal, 7-8.

we conducted here could be of great help; in fact, it could be crucial to make full sense of the submitted documents. Yet, our analysis focused on the identification of semantically different terms and we can only provide very vague insights into what these differences might be based and how they impact the stakeholders' opinions. Although we compared the most similar vectors corresponding to the word pairs of interest, this only gives us a rough idea of how the meanings might differ and still requires an inadmissibly high degree of *ad hoc* interpretation.

Therefore, and although this initial analysis already produced several notable takeaways for scholars and the Commission, the tool presented in this paper should be only the first building block of a fully-fledged tool for a more in-depth analysis.

CONCLUDING REMARKS

This paper set out to explore whether different stakeholders share a similar understanding of the many new challenges coming with the increased importance of digital markets. To do so, we contrasted the debate spurred by the consultation on the three EC inception impact assessments with the corresponding norms in the DSA and DMA proposals. We then employed computational tools to gain a fine-grained understanding of the stakeholders' feedback documents.

Analyzing replies to the EU Commission's public consultation, we find significant differences in stakeholders' use of central terms like for instance 'gatekeepers', 'procompetitive', 'collusion', and 'self-preferencing'.

While we believe that discerning latent differences in the use of certain terms is a crucial capability that could significantly enhance the consultation process, both lawmakers and legal scholars could benefit even further from quantitative text analysis if the tool we present in this paper is complemented by other NLP techniques.

For instance, 'topic modeling' could be used to get an intuitive understanding of which topics are the most relevant to stakeholders. Another powerful tool to be used more systematically in the future is sentiment analysis. While we used this technique to investigate the general attitude of each group of contributors towards certain words, one could use the same ide to cluster statements based on the sentiment of a group towards a certain concept or proposal to get a better understanding of how supporters and critics of a proposal are distributed and what their main concerns and arguments are.

If statements that are inputs to regulation are clustered based on document similarity measures, this could help to perceive certain similarities or

alliances between stakeholders, even across different groups like e.g., small companies and medium/large companies. As our analysis has shown, these clusters might not always look like what one would expect *prima facie*.

On a mor legal ground, computational tools could be used to trace back the influence of certain stakeholders by identifying those statements which are the most similar to the rules the Commission decided to propose. This could allow to gain a precise understanding of why rules were drafted in a certain way and greatly help the interpretation of norms in light of their *telos* and their drafting history.

Consequently, computational analysis could uncover novel insights into the provenance of a provision: which stakeholders asked for it, where does it come from? Given that the analysis of the drafting history and objectives of a norm is an essential part of its exegesis, these insights are essential for legal scholars. And especially when it comes to proposals as complex as the DMA and DSA, they could be of great value both for the Commissions and legal scholars, and hence something to look forward to.

Hence, that might suggest both to further the inquiry, and to make an effort to spread a common understanding of relevant terms.

* * *

Word count: 9526

APPENDIX

Annex 1. Results of preliminary data analysis to motivate corpus construction

Our clustering choice is based on two considerations: First, a qualitative analysis of the questionnaires accompanying the feedback documents[73] allowed us to get an understanding of which aggregation would cluster comparable feedback contributors together. Second, we conducted a quantitative analysis of the same questionnaires to ensure that our clusterization choices are solid. In particular, we sought to ensure that there is no statistically significant difference between *medium* and *large* entities in our sample since at least medium *companies* are often grouped with small, rather than large *companies*.

However, it needs to be noted that our feedback contributors are not only businesses, but also other types of organizations. This diversity could "smooth" the differences we would have expected to find if our sample included companies only. In fact, our qualitative analysis of the questionnaires suggested that medium entities in our sample are more comparable to large businesses/organizations.[74] To test the robustness of this perception, we analyzed the answers provided by medium and large entities to specific multiple choices questions.[75] We applied a Kolmogorov-Smirnov two-sample test[76] to understand if there is a statistically significant discrepancy between the distribution of the answers of the two groups. If that was the case, we would assume that theses answers must be considered as provided by two different populations, not allowing us to treat them as a unique cluster. The results of the test are shown in Figure 1:

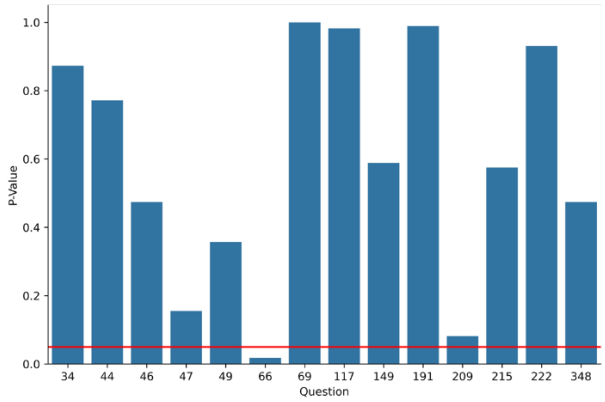


Figure 1: p-value distribution associated with the KS-two sample test applied to the answer distributions of the

[73] European Commission, Digital Services Act – deepening the internal market and clarifying responsibilities for digital services, 11 January 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers/public-consultation (accessed 28 January 2021).

[74] While this could be due to the idiosyncrasy of our sample, this finding also corresponds with scholarly literature. See e.g., Kemp, R., & Lutz, C. (2006). Perceived barriers to entry: Are there any differences between small, medium-sized and large companies *International Journal of Entrepreneurship and Small Business*, *3*(5), 538–553.

[75] We selected said questions based on what could be considered interesting for our research. A full list of the questions we selected can be found in the appendix.

[76] Hoboes, L. Jr. (1958). The significance probability of the Smirnov two-sample test. Matematica 3(5), 469-486.

considered questions. Red line highlights our significative tolerance value of p=0.05

Even using a very high tolerance **p-value** level of 0.05, only question no. 66[77] showed a statistically significant variation. This question alone however is mostly unrelated to our core research interest, and hence unlikely to compromise the validity of our clustering.

[77] *See* Annex 1.

Annex 2. A statistical test to identify semantic differences

We can model the relative distance d_t^{AB} of a word t in the corpus A and B be as:

$$d_t^{AB} = \gamma_t^{AB} + \mu_t^{AB} + u_t^{AB}$$

This takes into account a semantical term γ_t^{AB}, a non-semantical term (originated from the simple different words disposition in the two corpus) and a random term u_t^{AB}.

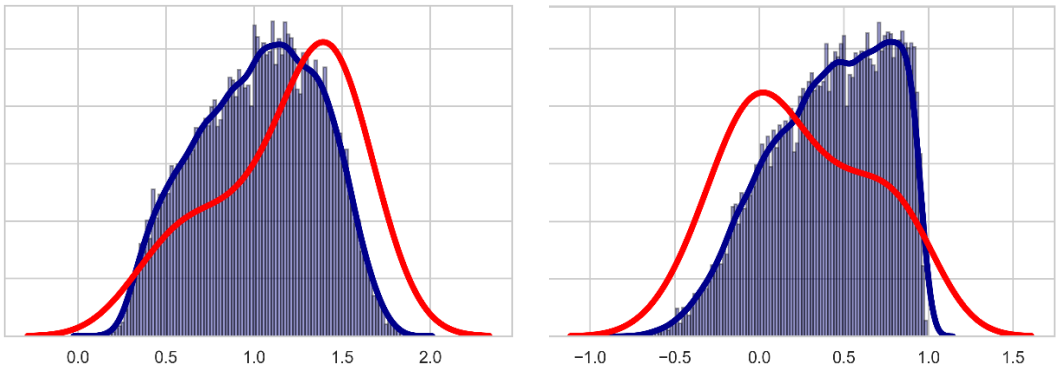
To isolate the semantic difference in the distances between words we found, we need to set two assumptions. Our first assumption is that words in the control vocabulary used for the Vector Space Alignment Transformation do not have a semantic difference, i.e., $\gamma_t^{AB} = 0$. This means we assume that stakeholders mean the same when they use words like 'and' or 'one'. Consequently, based on the distances between these 'control' words, we can construct an empirical distribution of the non-semantic distance between words. In this manner, we can get an idea of what a distance would look like if there was no semantic difference.

In a next step, we can hence compare the distance we observe for certain words with what we would expect if there was no semantic difference. Our p-value represents the probability to have a distance equal or greater than what we observe if our null-hypothesis, that there is no semantic difference between the same term in different corpuses, is actually true. If this probability is small enough, we can reject this null hypothesis with a small possibility of error. This is to say that the particular word has, indeed, a statistically significant semantic difference in the two corpuses. A general acceptance value for the p-value is 0.05, which we will use as the critical threshold for our analysis.

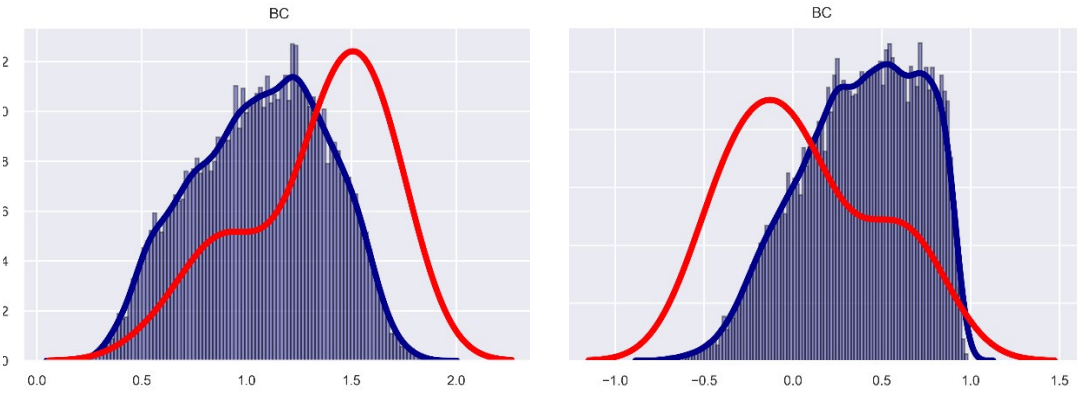
Annex 3 - Cumulative distribution of control distances

3.1 Corpuses AB - Cumulative distribution of control distances

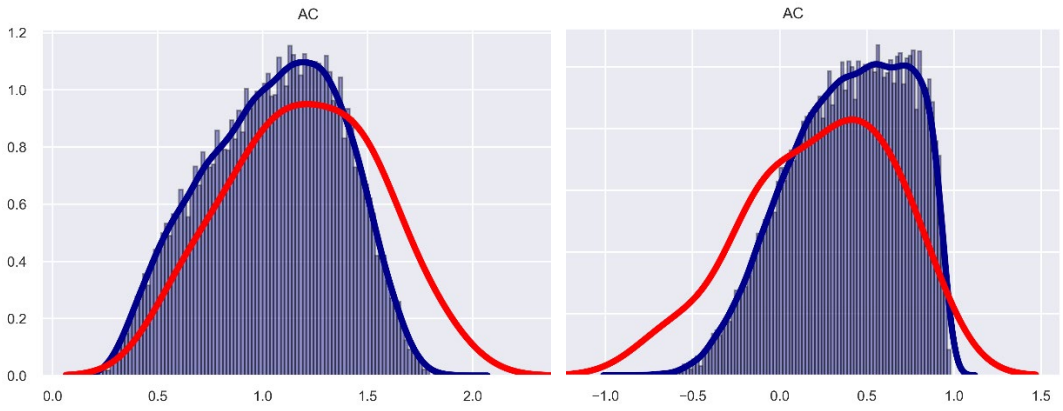
Figures 1 to 3 show the Cumulative Distribution of distances of control dictionary words (in blue) against cumulative distribution of distances of analyzed words (in red) for each corpus pair (i.e. corpus X against corpus Y).



3.2 Corpuses BC - Cumulative distribution of control distances



3.3 Corpus Pair AC - Cumulative distribution of control distances



Annex 4 - Questions manually selected from the DSA questionnaire78 to run the Kolmogorov-Smirnov two-sample test

Question 34: Did you ever come across illegal content online (for example illegal incitement to violence, hatred or discrimination on any protected grounds such as race, ethnicity, gender or sexual orientation; child sexual abuse material; terrorist propaganda; defamation; content infringing intellectual property rights, consumer law infringements)?
Question 44: Do you consider these measures([79]) appropriate?
Question 46: If your content or offering of goods and services was ever removed or blocked from an online platform, were you informed by the platform?
Question 47: Were you able to follow-up on the information?
Question 49: If you provided a notice to a digital service asking for the removal or disabling of access to such content or offering of goods or services, were you informed about the follow-up to the request?
Question 66: Does your organisation access any data or information from online platforms?
Question 69: Do you use WHOIS information about the registration of domain names and related information?
Question 117: What information would be, in your view, necessary and sufficient for users and third parties to send to an online platform in order to notify an illegal activity (sales of illegal goods, offering of services or sharing illegal content) conducted by a user of the service?

a) Precise location: e.g. URL
b) Precise reason why the activity is considered illegal
c) Description of the activity
d) Identity of the person or organisation sending the notification. Please explain under what conditions such information is necessary
e) Other, please specify

Question 149: In your view, is there a need for enhanced data sharing between online platforms and authorities, within the boundaries set by the General Data Protection Regulation? Please select the appropriate situations, in your view:
Question 191: Do you believe that in order to address any negative societal and economic effects of the gatekeeper role that large online platform companies exercise over whole platform ecosystems, there is a need to consider dedicated regulatory rules?
Question 209: Which, if any, of the following characteristics are relevant when considering the requirements for a potential regulatory authority overseeing the large online platform companies with the gatekeeper role:

a) Institutional cooperation with other authorities addressing related sectors – e.g. competition authorities, data protection authorities, financial services authorities, consumer protection authorities, cyber security, etc.
b) Pan-EU scope

[78] Note that the questions have been re-enumerated consecutively. The content remains the same.

[79] Question 43: 'What actions do online platforms take to minimise risks for consumers to be exposed to scams and other unfair practices (e.g. misleading advertising, exhortation to purchase made to children)?'.

c) Swift and effective cross-border cooperation and assistance across Member States
d) Capacity building within Member States
e) High level of technical capabilities including data processing, auditing capacities
f) Cooperation with extra-EU jurisdictions
g) Other

Question 215: Taking into consideration the parallel consultation on a proposal for a New Competition Tool focusing on addressing structural competition problems that prevent markets from functioning properly and tilt the level playing field in favour of only a few market players. Please rate the suitability of each option below to address market issues arising in online platforms ecosystems. Please rate the policy options below from 1 (not effective) to 5 (most effective).:

a) Current competition rules are enough to address issues raised in digital markets
b) There is a need for an additional regulatory framework Imposing obligations and prohibitions that are generally applicable to all large online platforms with gatekeeper power
c) There is a need for an additional regulatory framework allowing for the possibility to impose tailored remedies on individual large online platforms with gatekeeper power, on a case-by-case basis
d) There is a need for a New Competition Tool allowing to address structural risks and lack of competition in (digital) markets on a case-by-case basis.
e) There is a need for combination of two or more of the options 2 to 4.

Question 222: When you see an online ad, is it clear to you who has placed the advertisement online?

Question 348: In your view, is there a need to ensure similar supervision of digital services established outside of the EU that provide their services to EU users?