

The Influence of Self-Citations on Editorial Decisions and Impact Metrics: Description of an Agent-Based Model

Luiz Gabriel Correia
luiz.gabriel@ufabc.edu.br

January 2024

The model description follows the ODD (Overview, Design concepts, Details) protocol for describing individual- and agent-based models (Grimm et al. 2006), as updated by Grimm et al. (2019) [1].

1 Purpose and patterns

The purpose of this model is to explore the dynamics of the academic publishing process, particularly the strategies adopted by editors in the competition for better positions on the Journal Impact Factor (JIF) ranking. Specifically, the model aims to demonstrate that editor manipulation occurs in a zero-sum game scenario, as described by Wilhite et al., and predict which policies may be effective in removing incentives for practices that inflate impact metrics. To do so, the model is able to generate synthetic citation networks reproducing the patterns found in real-world citation networks.

Wilhite et al. shows that review articles, editorials, and coercive citations can be used to artificially inflate the JIF, and the influence of these strategies on indicators can be empirically verified, being highly correlated with the use of journal-level self-citations. Based on the observation of the high correlation between journal-level self-citations and the influence of the four manipulation strategies considered in the study, the authors propose the complete exclusion of self-citations from the JIF calculation as a possible policy to discourage the use of such manipulation strategies. According to the game theory model proposed by the authors, this policy would remove the incentive for manipulators while not harming legitimate actors. The model considers the three strategies addressed by the authors based on mechanisms used to increase self-citations: review articles, editorials, and coercive citations. The model aims to verify whether the complete exclusion of self-citations from the JIF calculation is sufficient to remove the incentives that would lead to the persistence or even growth of JIF manipulation.

There are a number of patterns used to calibrate and evaluate the model. Four distribution patterns and two growth patterns observed in the literature characterizing the dynamics of growth and structure of academic networks were selected, as described in the following list. The social networks generated by the model need to reproduce these patterns to be capable of generating networks with the same properties as those where competition between journals occurs.

1. Temporal evolution of the number of articles;
2. Temporal evolution of the number of journals;
3. Distribution of the number of documents per journal;
4. Distribution of the number of references per journal;
5. Distribution of the number of citations per journal;
6. Distribution of JIF scores per journal.

The growth patterns show a geometric progression, while the distribution patterns exhibit a typical power-law distribution generated from the so-called *Matthew Effect*.

To characterize the effects of editorial strategies on the JIF, we use four patterns observed in empirical data based on the observations of Wilhite et al., namely:

1. *Temporal evolution of the self-citation proportion*: the proportion of self-citations has been increasing over the years;
2. *Temporal evolution of the average JIF score*: the average Impact Factor score for all journals has also been increasing over time;

Variable name	Type/Units	Meaning
title	Text	Text identifier for the paper.
year	Integer number	The year of publication, counted from the start of the simulation.
doctype	Text	Type of publishable document. One of: “Article”, “Review” or “Letter”
publisher	Journal	Reference to the Journal in which the paper was published.

Table 1: State variables for *Papers*

3. *Pattern of interdisciplinary network clustering*: disciplines closer to each other tend to share more content, generating clusters of disciplines in more correlated areas;
4. *Pattern of specialization of manipulation strategies by disciplines*: qualitatively, we observe that the clustering of disciplines by prevalence of manipulation strategies is different for each strategy;
5. *Similarity between the patterns of clusters 3 and 4*: qualitatively, we see that the clusters generated by both patterns are similar, as expected by the theoretical model.

Initially, the evaluation of each pattern is qualitative. Once it is verified that the model is capable of qualitatively reproducing the selected patterns, a quantitative definition of each pattern is used to calibrate the parameters and find the configuration that best reproduces the data from the SJR sample and the findings from the literature.

Patterns 1 and 2 are explained by the theoretical zero-sum game model as a result of pressures to adopt manipulation strategies, and can be observed even when not considering the interdisciplinary network in the model. Patterns 3 to 5, on the other hand, emerge when considering the interdisciplinary network. The similarity between discipline clusters (pattern 5) is a strong prediction of the zero-sum game theoretical model, and its reproduction by the model is a good indicator of the *structural realism* of the proposed model.

2 Entities, state variables, and scales

The following entities constitute the model: agents representing elements of academic publishing (articles, journals, and disciplines) and links representing connections between the agents (citation relationship between articles, publishing relationship between articles/journals, shared content relationship between disciplines). The environment in which the agents reside is represented through a social network with the aforementioned connections, without explicitly representing spatial dimensions. Authors are not explicitly represented in this model, since it primarily addresses the dynamics among journals. Journals are represented as agents competing for article publications.

Articles represent any type of document publishable in an academic venue such as a journal or conference. The state variables of articles are presented in Table 1. Articles must have references to previously published articles and can be cited by future articles. These relationships are represented by undirected edges called *citation links*. *Journals* represent the venues where articles can be published. Journals have an editorial board that makes decisions about which articles to accept and what editorial strategy to follow; this decision-making process is abstracted in the model by treating the journal as a single decision-making agent. The state variables of journals are shown in Table 2. Articles published in a journal are connected by a directed edge called *publication link*, where the first vertex is a journal, and the second is an article. *Disciplines* represent sets of journals related by having publications on the same topics. In the model, disciplines determine the neighborhood of a journal, i.e., the journals for which editors can directly observe the strategies that other editors are following. The state variables of disciplines are presented in Table 3. Every journal is connected to one or more disciplines through a directed edge called *subject link*, where the first vertex is a discipline and the second is a journal. Disciplines are connected to each other through weighted undirected edges called *intersection links*, indicating the number of journals in common between the two disciplines.

As the model is based on connections between agents, there are no spatial dimensions. The temporal dimension must have an extension comparable to that of the chosen journal ranking for model calibration, the SCImago Journal Ranking (SJR). Longer periods can be used to attempt to reconstruct an advanced stage of manipulation strategy use from an initial point. The considered resolution is 1 year, with all articles published in the same year considered as published simultaneously. This resolution is sufficient for the study of common impact metrics such as the Impact Factor, as they are typically calculated over a window of 2 or 3 years.

Variable name	Type/Units	Meaning
name	Text	The name of the journal.
creation-year	Integer number	The foundation year of the journal counted from the start of the simulation.
JIF	Real number	Impact Factor for the journal calculated for the last 2 years.
results-estimates	Real vector	Vector of estimated probabilities for each possible ranking result and strategy.
strategy	Text	Which manipulation strategy the journal is using. One of the following: “Reviews”, “Editorials”, “Coerce” or “None”.
aversion	Real vector	Vector with 4 real coefficients between 0 and 1, each representing the journal’s degree of aversion to each strategy m .
sanction	Boolean	Whether the journal is currently under sanction.

Table 2: State variables for *Journals*

Variable name	Type/Units	Meaning
subject-area	Text	The subject area defining the discipline.

Table 3: State variables for *Disciplines*

Variable name	Type/Units	Meaning
citing-paper	Paper	The paper referencing the cited paper.
cited-paper	Paper	The paper being cited.

Table 4: State variables for *citation links*

Variable name	Type/Units	Meaning
published-paper	Paper	The paper being published.
publishing-journal	Journal	The journal publishing the paper.

Table 5: State variables for *publication links*

Variable name	Type/Units	Meaning
journal	Journal	The Journal being related to.
discipline	Discipline	The discipline the journal belongs to.

Table 6: State variables for *subject links*

Variable name	Type/Units	Meaning
discipline1	Discipline	One of the disciplines related to the link.
discipline2	Discipline	One of the disciplines related to the link.
numer-of-journals	Integer number, count	The number of journals in the intersection, i.e., belonging to both discipline1 and discipline2.

Table 7: State variables for *intersection links*

3 Process overview and scheduling

The dynamics of the model consist of 9 processes, with 5 processes executed by agents and 4 executed by the observer. The processes executed by agents are: *generate journals*, *generate papers*, *submit papers*, *editorial review*, and *update editorial strategy*. The processes executed by the observer are: *apply sanctions*, *update JIF*, *update view*, and *export data*.

The processes are executed as follows:

1. The observer executes *generate articles*, generating the quantity of articles to be published in the current year. Each article has the *year* variable updated to the current year.
2. The observer executes *generate journals*, generating the quantity of journals to be created in the current year. Journals differ in initialization regarding the disciplines they publish for, the manipulation strategy they use, and their aversion levels to manipulation practices. The process has three sub-actions, whose behavior depends on the input parameters:
 - (a) Create journals, where *creation-year* is initialized with the current year and *aversion* is initialized with random values;
 - (b) Choose one or more disciplines, creating a *subject link* for each;
 - (c) Assign an initial editorial strategy, initializing the *strategy* variable.
3. After the creation of articles and journals, each article will execute *submit articles*, which will choose a journal, among the newly created and old ones, to which it will be submitted for publication. The execution order is randomized. The criteria abstracting the authors' choice are described in Section 7.5. After the choice, a publication edge will be created between the article and the chosen journal.
4. After the submission of articles, each journal executes *editorial review* in randomized order. The process models the influence of the editorial board on the published articles. Each journal will decide, for each article submitted to it in the current year, what changes it will request for its publication to be accepted. Depending on the strategy, this process may change the *doc-type* variable of the article or alter the references of the submitted article, creating new citation edges to past articles.
5. After the publication of new articles, the observer will execute *update JIF*, which calculates the Impact Factor and updates the *JIF* variable for each of the journals, repositioning them in the ranking.
6. From the new ranking, each journal will execute *update editorial strategy* to decide whether to change its strategy based on what the neighborhood is doing and its own expectations of return. The execution order is also randomized. The process has the following sub-actions, detailed as submodels in Section 7.8:
 - (a) The journal calculates its neighborhood, i.e., the journals that publish in at least one common discipline;
 - (b) Based on the results obtained by its neighbors, the journal updates the probabilities for each possible outcome;
 - (c) The journal calculates the utility of each strategy and chooses the one that maximizes its expected utility.
7. After the journals have made their choices, the observer will execute *apply sanctions*, which will be executed for each journal in randomized order. For each journal using a manipulation strategy, there is a small chance of being caught, depending on the degree of manipulation and the strategy used. If the journal is caught, the applied sanction will be exclusion from the ranking for a given period. The submodel is detailed in Section 7.9.
8. At the end of each step, *update view* is executed by the observer and updates the screen and charts to enable the visualization of variables.
9. At the end of the simulation period, the process *export data* generates a ranking table as a csv file in the same format as the SJR.

The order of the processes follows the order of article publication: submission, editorial review, JIF calculation, and evaluation and decision-making by editors. As there is no interaction between agents during the creation of new articles and journals, the order in which they are created is irrelevant. However, in steps involving decision-making by agents (submission, article review, and update of editorial strategy), the order in which each agent executes the processes should be randomized to minimize the effect that actions performed first may have on the decisions of agents that execute last.

The submission and editorial review processes are modeled without considering the peer-review process and abstracting authors. This choice was made so that the model can focus mainly on reproducing two scale-free distribution patterns: the number of articles per journal and the number of citations per article. Considering authors in the model would imply also reproducing the distribution of articles per author and authors per journal. Additionally, as the purpose of the model is to reproduce journal competition for positions in the impact ranking, only factors necessary to reproduce empirically observed citation patterns are needed, so the peer-review process is not considered in this version of the model.

4 *Design concepts*

This model implements the zero-sum game model introduced by Wilhite et al. (2019) [2] as an agent-based model representing editorial decisions when faced with competition for impact metrics in journal rankings. The model replicates distributions observed in empirical data using accumulative advantage mechanisms shown by Watts and Gilbert (2011) [3], adapting the original approach of modeling articles and authors to modeling articles and journals.

The main emerging phenomenon generated by the model is the specialization pattern in the adoption of manipulation strategies among disciplines. It is expected that the pattern varies between disciplines, and disciplines with larger intersections are expected to share more dominant strategies. The exact distribution of each strategy among the journals, however, is an emergent phenomenon and should be highly sensitive to initial conditions. Scale-free network distributions of articles and citations are also emerging patterns generated from the accumulative advantage mechanisms implemented in the model, as observed in related models. The growth of scientific literature is determined directly by a geometric progression for the number of articles and for the number of journals. The agent decision-making process is based on a set of simple heuristics and on decision theory.

Adaptation of agents is modeled for articles and journals. Articles can be submitted to any journal. The decision criterion used by article authors can be abstracted as a random choice or based on a simple rule, where journals with higher Impact Factor are chosen with a higher probability than those with lower Impact Factor. This latter criterion can be considered an indirect goal search, with the goal of finding the best venue for publishing an article. Although a relatively big simplification of the choice process, the Impact Factor ranking is still considered a relevant aspect for choosing a publication venue [4]. A second possible adaptive behavior for articles is deciding whether the article will or will not comply to the requirements made during the editorial review. In case of a negative decision, the article may be resubmitted to another journal.

Journals have two adaptive behaviors: the first is editorial review, in which editors must decide whether they will require changes to the article in order to accept it for publication. The decision is also based on an indirect goal search, where journals will apply their manipulation strategy (if they see fit) in a proportion of cases determined by input parameters. Each of the three manipulation strategies imposes a different modification to the submitted article. The second adaptive behavior of journals is modeled as a direct goal search based on the game theory model. It models editorial decisions as a choice of which manipulation strategy to follow. The choice is made based on a utility function that attributes a score to each possible position gained in the Impact Factor ranking by the journal following a given strategy, contrasted with its aversion for using that strategy. The expected increase in ranking positions is calculated based on the observed changes in positions for neighboring journals and knowledge of the manipulation strategy they use. Journals keep in memory their probability estimates for each possible outcome in the ranking using each strategy, and update their estimates as new observations arise. Journals make their choice based on the principle of maximizing expected utility. Agent learning is not considered in this model beyond the heuristics and utility calculations described here.

The model assumes that information is asymmetric among editors. This means that journals are only aware of the manipulation strategies used by their peers, that is, by journals that publish in at least one discipline in common with themselves. All agents, journals, and articles have access to the overall JIF ranking. There is no uncertainty in agents' knowledge about the variables to which they have access.

The interaction between journals is represented by the exchange of information about the strategies used by journals, and can only occur among journals in the same discipline. It is a form of indirect interaction, since journals do not affect each other directly but have access to the strategies used by

others in their discipline. This asymmetry models the observation that the editorial board of a journal is generally shared among several journals in the same discipline, and academics in the same field tend to know and tolerate editorial practices that are normalized within their own field. The interaction of articles with journals occurs in submission and editorial review, while the interaction between articles occurs when one article cites another. Both interactions are direct, and when they occur, result in the creation of edges between agents. When created, an article can be submitted to any journal. As authors and content of articles are not explicitly modeled, articles are not classified by area or discipline until they are submitted to a journal. Articles can cite any other article that has been published in a journal. This assumption is reasonable if we consider that authors have access to an indexing database like the *Web of Science*. A journal interacts only with articles submitted to it.

Stochastic models are present in the initialization of the model, being used to initialize state variables as described in Section 3. The order in which agents perform actions is also stochastic, as well as the criterion for whether to apply certain heuristics or make certain choices, as described in Section 3. Stochastic elements are used in the model to represent processes not modeled by heuristics and to generate distributions with enough variation of observed patterns in the data.

The model makes use of one type of agent collectives: disciplines. Disciplines are aggregates of journals that publish articles on common themes, represented as a separate agent type with its own network of intersections with other disciplines. For the purposes of the model, disciplines determine the knowledge of journals about the manipulation strategies that other journals are using. One of the main patterns to be generated by this model is observed at the discipline level: the pattern of adoption of manipulation strategies among journals.

Relevant observations for the model include aggregated statistics that allow the visualization of the distribution of citations per article, articles per journals, the growth of articles and journals, the evolution of the proportion of self-citations among citations received by articles, the distribution of Impact Factor per journal, and the frequency of manipulative and non-manipulative journals in the rankings. The main output data generated by the model for statistical analysis is the JIF ranking, which can be compared with rankings available for empirical data and on which the same experiments conducted by [Willhite et al.](#) can be performed. Generated graphs, such as citation networks at the article and journal levels, and discipline intersection networks, etc., are also saved at the end of the simulation to be analyzed in comparison with graph metrics constructed with empirical data.

5 Initialization

In the initialization step, a set of disciplines is created, along with foundational papers (that have no references to previous papers) and foundational journals. How these agents are created depends on some initialization parameters. The foundational papers will be published on the foundational journals. The following actions are performed during initialization:

1. The observer initializes the variable *ticks* to correspond to the input parameter *start-year*.
2. The observer executes the process *generate disciplines* to generate disciplines and their relationships as a complete graph. The number of disciplines is fixed during the simulation, so this process is executed only once during initialization. The process has two sub-actions:
 - (a) Create discipline, initializing the variable *subject-area* with a randomized name or based on input data.
 - (b) Create intersection edges between all pairs of disciplines, initially with the weight *number-of-journals* set to zero.
3. The observer executes the process *generate papers* to generate a number of foundational papers determined by initialization parameters *foundational-papers*. Each article has the variable *year* updated to *start-year*.
4. The observer then executes the process *generate journals*, which generates a number of foundational journals given by the initialization parameter *foundational-journals*. Like other journals, foundational journals are created with three sub-actions:
 - (a) Create journals, *creation-year* is initialized with *start-year*, and the vector *aversion* is initialized with random values generated according to input parameters.
 - (b) Choose one or more disciplines, according to input parameters. For each discipline, an interdisciplinary edge will be created.

-
- (c) Assign an initial editorial strategy according to input parameters, initializing the variable *strategy*.
5. Each foundational paper will execute the process *submit papers* that will choose one of the foundational journals to which it will be submitted for publication. The execution order is randomized. All foundational papers make this choice using a uniform distribution, meaning they randomly choose one of the foundational journals. After the choice, a publication edge will be created between the paper and the chosen journal.

Initialization can be used to simulate the process of evolution and accumulation of academic production in a scenario with a few founding works, where the generic synthesis process described above can be executed based solely on the input parameters. The model can also be used to model the evolution of competition between journals from a specific year. In this second case, journals rankings for a specific year can be provided as input data. The expected format is a CSV file with the fields used by the SCImago Journal Ranking. SCImago ranking was chosen because it is publicly available for several years, including thousands of indexed journals based on the Scopus database, with their total publications and citations counts in the last year and in a 2 and 3 year window, as well as a field with subject areas in which each journal can be classified. A randomly sampled fraction of the journals present in the data for that year will be considered foundational journals and will be initialized with the name present in the data. The data on subject areas are used to reconstruct the network of disciplines with their intersections based on a specific scenario for one year. The foundational papers will be created following the same fraction as the number of documents published by each of the journals in the data, skipping the article submission step. However, the other variables are initialized by the same synthesis process, since the ranking data does not list journal publications. The fraction of input data that used as sample can be generated by a helper script.

In this model, we are concerned with the incentives that lead to the dominance of manipulation strategies among journal editors; therefore, in general, it is desirable to minimize the effect of specific initial conditions on the results. For this reason, sensitivity analyses to initial conditions are important for this model. We may also want to reconstruct the evolution of the dynamics between journals from a point in empirical data, and in this case, we are interested in understanding the possible scenarios that could occur starting from the same historical initial conditions but assuming that alternative policies had been implemented, e.g., whether self-citations are considered in the JIF calculation or not.

6 Input data

Similarly to the initialization, empirical data from the journal rankings can be used as input data in the model. In this case, the role of external data is to replace the geometric growth model imposed by an equation at each year with the growth of the number of articles and journals observed empirically in the data. For use as input, the dataset provided to the model needs to include the ranking for multiple years covering at least part of the temporal extent of the simulation. The SCImago ranking has data available for each year since 1999. The first year of the series will be used for initialization, and in subsequent years, new journals will be added as they appear in the rankings each year. If the data does not cover the entire temporal extent of the simulation, the direct growth function will be used for the years without data, with synthetic journals being added for each year without data.

The SCImago ranking can be obtained as a CSV file via the website for any year from 1999 onwards. However, to be used in the model, it is necessary to download multiple years of the ranking and consolidate them into a single CSV file that includes the *year* field, describing the year in which those data were published. As the ranking has thousands of indexed journals, it may also be necessary to take a sample of journals from each year. For each year, the sample needs to add a proportion of new journals while keeping the updated entries for journals present in the previous year. The sampling process can be carried out by a helper script before loading the data into the simulation.

Variable	Meaning
t	The current year simulated by the model.
G_D	The graph of discipline intersections. It's defined as a full graph with weighted edges.
D	The set of vertices of G_D , each representing a discipline.
$E(G_D)$	The edges of the graph G_D , each representing an intersection between disciplines.
$w(e), e \in E(G_D)$	The weight of an edge in $E(G_D)$ representing the number of journals with content in both disciplines.
G_J	The citation graph at journal level.
J	The set of vertices of G_J , each representing a journal.
$E(G_J)$	The set of edges of G_J , each oriented edge (a, b) representing a citation of journal b by journal a .
G_P	The citation graph at paper level.
P	The set of vertices of G_P , each representing a paper.
$E(G_P)$	The set of edges of G_P , each oriented edge (a, b) representing a citation of paper b by paper a .
H_P	The heterogeneous publication graph formed by the vertices $P \cup J$ to represent paper-publisher links.
$E(H_P)$	The set of edges of H_P , each edge (j, p) has a journal j that is the publisher of paper a .
H_D	The heterogeneous content graph formed by the vertices $D \cup J$ to represent the content relations between journals.
$E(H_D)$	The set of edges of H_D , each edge (d, j) has a journal j that publishes content on discipline d .

Table 8: Notation used to describe the variables and structures used in the model.

7 Submodels

7.1 Input parameters

The complete description of all input parameters is provided in Table 9. Table 8 explains the notation used to represent variables and structures necessary to formalize the model.

7.2 Geometric growth of scholarly data

The growth of scientific literature observed in empirical data over the past decades is exponential, as evidenced by patterns in the number of published articles, new authors, and new journals. The number of active journals in the SCIMago ranking exhibited slower growth from 1999 to 2018 when compared to the number of documents. Both growths can be modeled by a geometric progression in different parameters. As new articles and journals need to be added over time in the simulation, the number of articles created at each step is given by the following expression:

$$P_{t+1} = r_P P_t \quad (1)$$

Where P_t is the number of articles created in year t , and r_P is a constant input parameter determining the growth rate of P . If $r_P > 1$, the number of articles will grow exponentially as t increases. On the other hand, the growth of the number of journals is determined by the input parameter r_J , representing the number of journals to be added each year in the simulation. The parameters r_J and r_P can be estimated based on empirical data.

The equation for journals is quite similar, but in this case, the growth rate r_J will have a smaller value compared to r_P .

$$J_{t+1} = r_J J_t \quad (2)$$

This type of modeling used to represent the growth of journals and articles has been proved enough to replicate the scale-free distribution patterns needed to capture the dynamics of citation metrics [3]. The form of a geometric progression was chosen instead of an explicit exponential expression because

Parameter	Text name	Type/Domain	Meaning	Used by submodel
$t_{\max}$	years	Integer number, $[0, \infty)$	The total number of years to be simulated by the model.	Environment
s	seed	Integer number	The seed value used by the pseudorandom number generator. It needs to be saved in order to reproduce the experiments.	Environment
$ D $	number of disciplines	Integer number, $[1, \infty)$	The total number of disciplines to be created in the initialization.	Generate disciplines
r_J	journal growth	Real number, $[0, \infty)$	The growth rate of the number of journals created each year.	Generate journals
r_P	paper growth	Real number, $[0, \infty)$	The growth rate of the number of papers created each year.	Generate papers
$ P_0 $	foundational papers	Integer number, $[1, \infty)$	The number of foundational papers to be created at initialization.	Generate papers
$ J_0 $	foundational journals	Integer number, $[1, \infty)$	The number of foundational journals to be created at initialization.	Generate journals
$\delta^+(G_P)$	references per paper	Integer number, $[1, \infty)$	The mean number of references a paper has. In graph notation, it's the mean out-degree in the paper citation network G_P .	Generate references
α_{age}	recency weibull param 1	Real number, $[1, \infty)$	The scale parameter of the Weibull Cumulative Distribution Function (CDF), also called λ . Used for defining the recency of a paper.	Choose paper
β_{age}	recency weibull param 2	Real number, $[1, \infty)$	The scale parameter of the Weibull Cumulative Distribution Function (CDF), also called k . With $\beta > 1$ the CDF represents aging.	Choose paper
α_{prestige}	recency weibull param 1	Real number, $[1, \infty)$	The scale parameter of the Weibull Cumulative Distribution Function (CDF), also called λ . Used to define the prestige of a journal.	Submit paper
β_{prestige}	recency weibull param 2	Real number, $[1, \infty)$	The scale parameter of the Weibull Cumulative Distribution Function (CDF) used to define the prestige of a journal.	Submit paper
$P(S = \text{"copy publisher"})$	chance copy publisher	Real number, $[0, 1]$	The probability a paper being submitted will choose the publisher of a recent paper as its journal for submission. This simple choice reproduces the cumulative advantage effect.	Choose journal
$P(S = \text{"copy ref"})$	chance copy reference	Real number, $[0, 1]$	The probability a paper will choose its references by copying them from another paper chosen with the strategy S_{ref} .	Generate references
S_{ref}	reference strategy	{ "Copy from a recent paper", "Copy from any paper", "Any recent paper", "Any paper" }	The strategy used to choose a reference for a paper.	Generate references
S_{submit}	submit strategy	{ "Copy from recent paper", "Copy from cited paper", "Top journal", "Any journal" }	The strategy used by a paper to choose a journal to be its publisher.	Choose journal
S_{global}	global editorial strategy	{ "None", "Coerse", "Reviews", "Editorials", "Utility" }	The manipulation strategy used by all journals to inflate the JIF. All options except "Utility" are fixed strategies among the journals.	Editorial review
r_{start}	start manipulator ratio	Real number, $[0, 1]$	The initial ratio of journals using a manipulation strategy. Has no effect if <i>editorial strategy</i> is "None", if it's "Utility" the ratio can change with time since journals can switch strategies.	Generate journals
r_{coersed}	coersed citations ratio	Real number, $[0, 1]$	The proportion of references of a submitted paper that will be citations coerced by the editor during the review.	Coerse citations
r_{reviews}	reviews ratio	Real number, $[0, 1]$	The proportion of self-serving review articles a journal will try to publish per year as part of its manipulation strategy.	Publish review
$r_{\text{editorials}}$	editorials ratio	Real number, $[0, 1]$	The proportion of editorials in the form of "Letters" the journal will try to publish per year as part of its manipulation strategy.	Publish editorial
$r_{\text{detection}}$	manipulation detection rate	Real number, $[0, 1]$	The max rate at which journal manipulation can be detected and sanctioned in a given year of the simulation.	Apply Sanction
r_{sjr}	ranking sample ratio	Real number, $[0, 1]$	The ratio of entries to be randomly sampled from the SJR ranking when using as input data.	
t_{sanction}	sanction duration	Integer number, $[0, t_{\max}]$	The number of years for which the journal will be excluded from the ranking.	Apply Sanction
u_{reviews}	references review factor	Real number, $[1, \text{inf})$	A multiplication factor that increases the number of references reviews can have from the base parameter $\delta^+(G_P)$.	Editorial review
$u_{\text{editorials}}$	references ediorial factor	Real number, $(0, 1]$	A multiplication factor that decreases the number of references editorials can have from the base parameter $\delta^+(G_P)$.	Editorial review
t_s	strategy start year	Integer number, $(1, t_{\max})$	The simulated year in which the journals start applying their strategies, useful to compare the impact on citations and self-citations.	Editorial review
$a_{\max}$	max aversion	Real number, $[0, 1]$	Sets the max value of the journal variable <i>aversion</i> , the adoption ratio a journal needs to see among his peers to also adopt a manipulation strategy.	Update editorial strategy
$d_{\max}(J)$	max disciplines per journal	Integer number, $[1, D]$	The max number of disciplines a journal will publish content to. The higher the number, the higher the intersections between disciplines.	Update editorial strategy
$P(S = \text{"copy discipline"})$	chance copy discipline	Real number, $[0, 1]$	The probability of a new journal copying the disciplines of an existing journal in his main discipline when defining its own subject areas.	Choose discipline

Table 9: Input parameters and their descriptions.

representing only the growth rate of *articles/year* or *journals/year* facilitates its substitution with empirical input data for specific years. Similarly, in the absence of more data, these can be substituted by the growth determined by the corresponding geometric or arithmetic progressions.

7.3 Generate disciplines

This submodel generates a network of disciplines as a complete graph, where each vertex is an agent representing a discipline, and each edge represents the intersection between a pair of disciplines regarding the journals publishing in both subject areas. If A is the set of journals publishing in discipline A and B is the set of journals publishing in discipline B, then the weight of the edge $A \cup B$ corresponds to the number of journals in the intersection of A and B , $|A \cap B|$. Initially, all weights are initialized to 0 and are updated during the creation of journals.

The number of disciplines in the network is fixed throughout the simulation. This simplification is reasonable since, by observing the SCImago ranking over the last 20 years, there hasn't been a significant change in the number of categories considered by the ranking, as opposed to the number of journals or publications.

The *generate disciplines* procedure used to synthetically create the discipline network is presented in Algorithm 1.

Algorithm 1 Generate disciplines

```

procedure GENERATEDISCIPLINES( $n$ )
   $D \leftarrow \{\}$ 
  for  $i \leftarrow 1, n$  do
    name  $\leftarrow$  "Discipline" +  $i$                                  $\triangleright$  Generate a unique name to discipline
     $d \leftarrow \text{CREATEAGENT}(\text{Discipline})$                      $\triangleright$  Creates a new agent of type "Discipline"
     $d.\text{name} \leftarrow \text{name}$ 
     $D \leftarrow D \cup \{d\}$ 
  end for
  for all  $(a, b) \in D^2$  with  $a \neq b$  do
     $e \leftarrow \text{CREATEEDGE}(a, b)$                                  $\triangleright$  Creates an edge, if it doesn't exist
     $e.w \leftarrow 0$                                                $\triangleright$  Initializes the weight to 0
  end for
end procedure

```

The procedure *generate disciplines* is only executed when no external input data is used, as it creates the disciplines before the creation of journals. When external input data is provided, disciplines will also be created as new journals are added. A new discipline is created whenever a journal from an area not yet present in the model is added.

7.4 Generate journals

New journals will be generated at each simulation step, therefore, this submodel will be executed multiple times. Initially, journals are created as part of the content network H_D , a heterogeneous graph formed by disciplines and journals, with edges connecting journals to the disciplines in which they will publish content. The goal of this model is to iteratively construct the graph H_D in such a way as to enable the calculation of edge weights in the discipline graph G_D . No publications such as articles are created at this stage. The pseudocode is shown in Algorithm 2.

The procedure *generate journals* is used to generate synthetic journals. To generate journals from SCImago ranking input data, the procedure *generate SJR journals* (Algorithm 3) should be used.

7.4.1 Choose discipline

When journals are synthetically generated, it is necessary to specify a method for choosing the disciplines for which it will have publications. This submodule implements a simple strategy for making this choice during the creation of a journal. The procedure begins by choosing a primary discipline from the discipline set with a uniform distribution. From this primary discipline, one or more secondary disciplines can be chosen. The method for choosing secondary disciplines is similar to the accumulative advantage mechanism used for citations: with a fixed probability p_r , new journals choose one of the disciplines that other journals in their primary discipline also publish. With probability $1 - p_r$, the secondary discipline is chosen uniformly. This simple criterion allows generating networks of disciplines with different degrees of proximity, forming clusters of more related journals than others, as observed in

Algorithm 2 Generate journals

```
procedure GENERATEJOURNALS( $n$ )
   $J \leftarrow \{\}$ 
  for  $i \leftarrow 1, n$  do
    name  $\leftarrow$  "Journal" +  $i$                                  $\triangleright$  Generates a unique name to journal
     $j \leftarrow$  CREATEAGENT(Journal)                             $\triangleright$  Creates a new agent of type "Journal"
     $j.name \leftarrow$  name
     $j.aversion \leftarrow a \in_U [0..a_{\max}]$ 
     $d \leftarrow$  CHOOSEJOURNAL( $j$ )                                 $\triangleright$  Choose disciplines for  $j$  (see submodel 7.4.1)
    SETUPINITIALSTRATEGY( $j$ )                                     $\triangleright$  Setup initial editorial strategy (see submodel 7.4.2)
     $J \leftarrow J \cup \{j\}$ 
  end for
end procedure
```

Algorithm 3 Generate journals from input data

```
procedure GENERATEJOURNALSJR(year)
  ranking  $\leftarrow$  SAMPLERANKINGSJR(year,  $r_{\text{sjr}}$ )               $\triangleright$  Reads a sample from the SJR following the
  proportion  $r_{\text{sjr}}$ 
   $J \leftarrow \{\}$ 
  for entry in ranking do
    if there is no  $j \in J$  with  $j.name = \text{entry.name}$  then
       $j \leftarrow$  CREATEAGENT(Journal)                             $\triangleright$  Creates a new agent of type "Journal"
       $j.name \leftarrow \text{entry.name}$ 
       $j.aversion \leftarrow a \in_U [0..a_{\max}]$ 
      for each subject-area in entry["Categories"] do
        if there is no  $d \in D$  with  $d.name = \text{subject-area}$  then
           $d \leftarrow$  CREATEAGENT(Discipline)                     $\triangleright$  Creates a new agent of type "Discipline"
           $d.name \leftarrow \text{subject-area}$ 
        end if
        Let  $d \in D$  with  $d.name = \text{subject-area}$ 
         $e \leftarrow$  CREATEEDGE( $j, d$ )                             $\triangleright$  Creates an edge between journal  $j$  and discipline  $d$ 
      end for
      SETUPINITIALSTRATEGY( $j$ )                                     $\triangleright$  Setup initial editorial strategy (see submodel 7.4.2)
       $J \leftarrow J \cup \{j\}$ 
    end if
  end for
end procedure
```

empirical data. The pseudocode for choosing a discipline d from a primary discipline d_0 is detailed in Algorithm 4.

Algorithm 4 Choose secondary discipline from d_0

```

procedure CHOOSEDISCIPLINE( $j$ )
  Let  $d_0 \in_U D$  ▷ Choose a primary discipline  $d_0$  with uniform distribution
  CREATEEDGE( $j, d_0$ ) ▷ Creates an edge between journal  $j$  and discipline  $d_0$ 
  Let  $m \in_U [1..D_{\max}]$  ▷ Choose the number of disciplines  $m$  for  $j$  in the given range
  repeat
    Let  $x \in_U [0..1]$ 
    if  $x < P(S = \text{"copy discipline"})$  then
      Copy a discipline from other journal with given probability
      Let  $j' \in_U N_{H_D}(d_0)$  with  $j' : N_{H_D}(j') > 1$ 
      Let  $d \in_U N_{H_D}(j')$  with  $d \neq d_0$ 
    else
      Seja  $d \in_U D$  with  $d \neq d_0$ 
    end if
     $e \leftarrow \text{CREATEEDGE}(j, d)$  ▷ Creates an edge between journal  $j$  and discipline  $d$ 
  until  $\delta_{H_D}(j) = m$  ▷ Repeat until the number of subjects in  $j$  is  $m$ 
end procedure

```

7.4.2 Setup initial strategy

The Algorithm 5 assigns an initial strategy to each journal based on the input parameters.

Algorithm 5 Setup initial editorial strategy

```

procedure SETUPINITIALSTRATEGY( $j$ )
   $j.\text{strategy} \leftarrow \text{"None"}$ 
  Let  $x \in_U [0..1]$ 
  if  $x < r_{\text{start}}$  then
    if  $S_{\text{global}} \neq \text{"Utility"}$  then
       $j.\text{strategy} \leftarrow S_{\text{global}}$ 
    else
      Let  $s \in_U \{\text{"Coerse"}, \text{"Reviews"}, \text{"Editorials"}\}$ 
       $j.\text{strategy} \leftarrow s$ 
    end if
  end if
end procedure

```

7.5 Submit papers

This submodule abstracts the authors' choice of which journal to publish in. Since the content of papers and their authors is not represented in the model, this choice does not consider the subject areas of each journal, thus, it is assumed that the discipline of a paper is labeled based on the journal where it was published. The choice can be based on four criteria, defined by the parameter S_{submit} :

- "Any journal": Randomly chooses any journal, new or old, with a uniform distribution. This criterion does not reproduce any accumulative advantage effect and is useful as a control in experiments.
- "Top journal": Randomly chooses a journal, but with a distribution reflecting the prestige of each journal based on its position in the JIF ranking. The lower the position, the lower the probability of the journal being chosen. This criterion models the accumulative advantage of the number of articles per journal as depending on the journal's position in the Impact Factor ranking. A Weibull distribution is used analogously to the modeling of the age of articles, with input parameters α_{prestige} and β_{prestige} for the cumulative distribution function (CDF).
- "Copy from recent paper": With probability $P(S = \text{copy publisher})$, it first chooses a recent article with the same criteria used for references (see Section 7.6) and then chooses the journal

that published this article. The rest of the time, it chooses a journal randomly to ensure that new journals are chosen. This criterion copies the choice that other articles are making, thus providing a simple mechanism that reproduces the accumulative advantage effect on the number of articles per journal.

- “Copy from cited paper”: With probability $P(S = \text{copy publisher})$, it first chooses a recent and cited article with the same criteria used for references (see Section 7.6) and then chooses the journal that published this article. The rest of the time, it chooses a journal randomly to ensure that new journals are chosen. This criterion copies the choice of the most cited articles, thus reproducing the accumulative advantage effect on the number of articles per journal, indirectly depending on both the number of articles published and the number of citations of these articles.

The pseudocode implementing these criteria is presented in Algorithm 6. The procedure applies the criterion to all articles published in the current step of the simulation.

Algorithm 6 Strategies for paper submission

```

procedure SUBMITPAPERS
  for all  $p \in P$  do
     $j \leftarrow \text{CHOOSEJOURNAL}(S_{\text{submit}})$  ▷ Choose  $j$  using input parameter  $S_{\text{submit}}$ 
     $\text{CREATEEDGE}(p, j)$ 
  end for
end procedure
function CHOOSEJOURNAL( $S_{\text{submit}}$ )
   $x \in_U [0..1]$ 
  if  $x < P(S = \text{“copy publisher”})$  then ▷ Choose  $j$  following  $S_{\text{submit}}$ 
    if  $S_{\text{submit}} = \text{“Top journal”}$  then
       $\text{Seja } j \in_W J$  ▷  $W$  is a Weibull CDF representing the the prestige of  $j$ 
    else if  $S_{\text{submit}} = \text{“Copy from recent paper”}$  then
       $p \leftarrow \text{RECENTPAPER}()$ 
       $j \leftarrow p.\text{publisher}$ 
    else if  $S_{\text{submit}} = \text{“Copy from cited paper”}$  then
       $p \leftarrow \text{CITEDRECENTPAPER}()$ 
       $j \leftarrow p.\text{publisher}$ 
    end if
  else ▷ Choose  $j$  with uniform distribution
     $\text{Let } j \in_U J$ 
  end if return  $j$ 
end function

```

7.6 Editorial review

This submodule represents the influence of editors on the citations flows and content of papers, being the mechanism by which editorial decisions turn into manipulation of self-citations. Of the three manipulative strategies, two (reviews and editorials) involve changing the *doctype*, representing articles of a specific type requested by the editors, and the other (coercive citations) involves inserting citations to other papers. Journals begin to apply their manipulation strategies from the year t_S defined as an input parameter. The behavior of each strategy is discussed in the subsections below.

7.6.1 Review Articles

The article will be converted into a retrospective review article with probability r_{reviews} , reviews have on average a greater number of references when compared to a regular article. The number of references is determined by the input parameter $\nu_{\text{reviews}} > 1$, a multiplicative factor that multiplies the average number of references $\bar{\delta}^+(G_P)$ for all papers. As the model represents retrospective reviews, the selected references will be self-citations, i.e., they will always be chosen from papers from the same journal.

Individually, this is the strategy that concentrates the most on self-citations. Its disadvantage is that reviews correspond to a relatively small proportion of the total number of articles (thus the value of r_{reviews} should be small in the model for the mechanism to be realistic).

7.6.2 Editorials

The article will be converted into an editorial with probability $r_{\text{editorials}}$. Editorials will have fewer references than research articles, but like reviews, they will all be self-citations. The number of references is determined by the input parameter $\nu_{\text{editorials}} < 1$, a factor multiplying the average number of references $\bar{\delta}^+(G_P)$. Editorials should also correspond to a relatively small proportion of the total number of articles. The advantage of editorials over review articles lies in a peculiarity of the Impact Factor calculation, in which citations received by an editorial are included in the numerator, but since it is not by itself considered a citable item, it does not alter the denominator count.

7.6.3 Coercing citations

In this strategy, some self-citations will be added to the set of citations of the article at the request of the editors. With this strategy, fewer self-citations can be added per article; however, there is no limitation on the number of articles that can have citations added by coercion, unlike the other two strategies. The exact number of self-citations added will be a fraction of the total original citations, given by the parameter r_{coersed} .

7.7 Update JIF

The Impact Factor calculation is done with a 2-year window, being updated after the second year of the simulation ($t > 2$). The calculation follows the peculiarity pointed out in [Wilhite et al.](#). Only citable documents (i.e., articles and reviews) are considered in the denominator, while citations to all documents, including letters and editorials, are counted in the numerator. The pseudocode is provided in Algorithm 7. The procedure *update ranking* is detailed in Section 7.10.

Algorithm 7 Update Journal Impact Factors

```

procedure UPDATEJIF
  if  $t > 2$  then
    for all  $j \in J$  do
      Let  $P_2(j)$  all documents published in years  $t - 1$  e  $t - 2$  on  $j$ 
      Let  $P_2^*(j)$  all citable documents in  $P_2(j)$ 
      if  $|P_2^*(j)| > 0$  then
         $\sigma \leftarrow \sum_{p \in P_2^*(j)} \bar{\delta}_{G_P}^-(p)$   $\triangleright \sigma$  is the sum of all citations received by  $j$ 
         $j.\text{JIF} \leftarrow \frac{\sigma}{|P_2^*(j)|}$ 
      else
         $j.\text{JIF} \leftarrow 0$ 
      end if
    end for
    UPDATERANKING( )
  end if
end procedure

```

7.8 Update editorial strategy

The submodel defines how the editorial board of journals decides whether a manipulation strategy will be followed in the next simulation step. This submodel is the mechanism that generates the main phenomena to be studied in this model, being a direct adaptation of the zero-sum game model proposed by [Wilhite et al.](#). The model is based on the hypothesis of utility-maximizing rationality. To calculate the expected utility for each of the possible outcomes in the ranking, a utility function is defined for each journal along with a Bayesian estimator that allows them to have their own expectations about these results based on the asymmetric information that journals have access regarding the success of different strategies.

7.8.1 Calculate neighborhood

The neighborhood of a journal j is the set of all journals that have publications in any of the disciplines in which j publishes. Formally, let $N_{H_D}(j) \subseteq D$ be the set of neighbors of j in the graph H_D , and let $N_{H_D}(d) \subseteq J$ be the set of journals neighboring the discipline d in H_D . The set of pairs or neighbors of journal j is denoted $\mathcal{N}(j)$ and is defined as:

$$\mathcal{N}(j) = \bigcup_{d \in N_{H_D}(j)} N_{H_D}(d) \quad (3)$$

7.8.2 Utility function for journals

Journals can choose one of four distinct strategies: using reviews, editorials, coercive citations, or using none. The model will execute different behaviors for each strategy, either successfully inflating the JIF or punishing manipulative agents. The SCIMago ranking classifies journals into quartiles Q1, Q2, Q3, or Q4 based on their positions in the ranking [5]. This classification is used in the model to define the utility functions for journals. The utility of j is defined based on the trend of j 's position change in the ranking and the change between quartiles. Thus, the number of possible outcomes is limited to the combination of quartiles with three categories of trends (increase, stability and decrease).

Each journal j has a utility function $U_j(r, m)$ that assigns a score to each possible ranking result r given the strategy m used. $U_j(r, m)$ also depends on the aversion degree that j gives to manipulation practices. Assuming that each journal's sole objective is to maximize its position in the Impact Factor ranking, a simple expression defining its utility based on these parameters is given by:

$$U_j(r, m) = \alpha(1 - a_m)r - \beta a_m |r| \quad (4)$$

Here, $b_j = \alpha(1 - a_m)r$ represents gains, which are proportional to the result obtained in the ranking. On the other hand, $c_j = \beta a_m |r|$ represents costs as being proportional to the change caused by the strategy in the ranking, i.e., the absolute number of positions changed $|r|$ adjusted by a_m , the aversion coefficient of j to strategy m .

The utility function is linear with subtractive costs. This choice is based on the assumption that, despite gains being proportional to the gain of positions in the ranking, any ranking change that can be attributed to a manipulation strategy must have a negative component proportional to the agent's aversion.

If $a_m = 0$, the expression reduces to $U_j(r, m) = \alpha r$ and depends only on the result obtained in the ranking, being indifferent to the strategy used. If $a_m = 1$, the expression reduces to $U_j(r, m) = -\beta |r|$ and any ranking change resulting from m will have negative utility. Intermediate values of a_m changes the degree to which the journal values increases in the ranking using strategy m . The submodel parameters α and β allow adjusting the shape of the function for each term independently.

Possible ranking outcomes are given by a function $r(\Delta, \Delta_Q)$ defined in terms of the change in the journal's position (Δ) and its respective quartile change (Δ_Q) in the current ranking with respect to the previous one. The definition of $r(\Delta, \Delta_Q)$ is given below.

$$r(\Delta, \Delta_Q) = \begin{cases} \Delta_Q & \text{if } \Delta = 0 \\ \Delta_Q + 1 & \text{if } \Delta > 0 \\ \Delta_Q - 1 & \text{if } \Delta < 0 \end{cases}$$

The change between quartiles can vary between -4 (if a journal from Q1 is removed from the ranking) to $+4$ (if the journal enters the ranking directly in Q1). The change between quartiles is defined as $\Delta_Q = Q(p_i) - Q(p_{i-1})$, where $Q(p_i)$ is the quartile to which the position p belongs in the ranking of year i . The function $Q(p)$ is shown in Algorithm 8. We have that if $\Delta_Q > 0$, then $\Delta > 0$, and if $\Delta_Q < 0$, then $\Delta < 0$. Considering these constraints, there are a total of 11 distinct possible outcomes for r .

7.8.3 Editorial decisions maximizing expected utility

For each possible result in the ranking, the agents estimate the probability $P(r|m)$ of such a result r being obtained given that the strategy m was used. This estimate is made based on previous observations of their peers' results and knowledge of the strategy they used. Agents choose which strategy they will follow by calculating the expected utility for each strategy based on their probability estimates (which depend on the strategies used by peers) and choosing the one that maximizes expected utility.

The results obtained by each journal can be calculated from the second ranking. Initially, the probability *a priori* $P(r|m)$ is initialized uniformly among the total possible outcomes. In subsequent steps, the probability $P(\delta|m)$ is updated with the probability estimated in the previous iteration $t - 1$ being used as *prior* in the new estimate:

$$P(r|m) = \frac{P(m|r)P(r)}{P(m)} = \frac{P(m|r)P_{t-1}(r|m)}{P(m)} \quad (5)$$

Algorithm 8 Calculate the quartile of a position p in the ranking

```

 $Q_1 \leftarrow \text{round } \frac{1}{4}(|J| + 1)$ 
if  $|J|$  is pair then
     $Q_2 \leftarrow \frac{|J|}{2}$ 
else
     $Q_2 \leftarrow \frac{|J|+1}{2}$ 
end if
 $Q_3 \leftarrow \text{round } \frac{3}{4}(|J| + 1)$ 

function  $Q(p)$ 
    if  $p \leq Q_1$  then return 1
    else if  $p > Q_1$  and  $\leq Q_2$  then return 2
    else if  $p > Q_2$  and  $\leq Q_3$  then return 3
    else if  $p > Q_3$  then return 4
    end if
end function

```

The *likelihood* $P(m|r)$ is calculated from the observed data with the expression:

$$P(m|r) = \frac{|\mathcal{N}_{(r,m)}(j)|}{|\mathcal{N}_r(j)|} \quad (6)$$

Where $\mathcal{N}_r(j) \subseteq \mathcal{N}(j)$ is the set of neighbors of j that obtained result r in the ranking, while $\mathcal{N}_{(r,m)}(j) \subseteq \mathcal{N}_r(j)$ is the set of neighbors adopting strategy m that obtained result r in the ranking. If $\mathcal{N}_r(j) = 0$ the *likelihood* will be estimated as a uniform distribution among the 4 strategies.

The normalization term $P(m)$ is defined by:

$$P(m) = \sum_r P(m|r)P_{t-1}(r|m) \quad (7)$$

Once the probabilities are estimated, each agent calculates the expected utility of each strategy:

$$E[U_j(m)] = \sum_r P(r|m)U_j(r, m) \quad (8)$$

Finally, each agent updates its strategy to the strategy m for which $E[U_j(m)]$ is maximum:

$$j.\text{strategy} \leftarrow \arg \max_m E[U_j(m)] \quad (9)$$

7.9 Apply sanctions

This submodel consists of a simple heuristic that represents one of the forms of sanction that journals accused of misconduct may face: being excluded from the journal ranking. In the context of this model, where the agents' utility is directly determined by their position in the ranking, being excluded from the ranking is the same as being placed below the last position, resulting in the worst outcome evaluated by the agents' utility function (see Section 7.8).

The decision to exclude a journal from the ranking depends on whether the misconduct is detected or not. In this model, this behavior is modeled based on a base parameter $r_{\text{detection}}$, the maximum rate at which manipulative journals can be detected and penalized in a given year of the simulation. This base rate is modified by factors that depend on the journal and the strategy in question. Each journal will have a probability $\sigma(\Delta, m)$ of being detected and subjected to the sanction. The probability $\sigma(\Delta, m)$ is defined as a fraction of $r_{\text{detection}}$ and depends on the number of positions gained by the journal and the degree of adoption of strategy m in the discipline. The idea behind this assumption is to have a higher level of suspicion directed towards very large position gains. On the other hand, the more common a manipulation practice becomes in a discipline, the less likely it is for this behavior to be considered potential misconduct.

A simple expression for σ that models the above assumptions for a discipline d is:

$$\sigma_d(\Delta, m) = \left(1 - \frac{|N_d(m)|}{|N_d|}\right) (1 - e^{-\Delta})r_{\text{detection}} \quad (10)$$

Variable	Meaning and Usage
$ J_t $	The size of the set of journals at the year t . Used to plot the number of journals per year.
$ P_t $	The size of the set of papers at the year t . Used to plot the number of papers per year.
$ E(G_P(t)) $	The total number of citations at the year t . Used to plot the number of citations per year.
$\delta_{G_P}^-(p)$	The number of citations received by paper $p \in P$. Used for a histogram plot of citations per paper.
$\delta_{G_J}^-(j)$	The number of citations received by journal $j \in J$. Used for a histogram plot of citations per journal.
$\delta^*(j) = \{(j, j') : (j, j') \in \delta_{G_J}^-(j) \text{ e } j = j'\} $	The number of self-citations at the journal level in the year t . Used to plot the relative evolution of self-citations per year, normalizing $\delta^*(j)$ by $\delta_{G_J}^-(j)$.
$\delta_{H_J}(j)$	The number of papers published by journal $j \in J$. Used for a histogram plot of papers per journal.
$j.\text{JIF}(t)$	The JIF of journal j in year t . Used for a histogram plot of journals per JIF and to plot the distribution of JIF vs Total Docs.
$\bar{\text{JIF}}(m)$	The average JIF among the journals adopting editorial strategy m . Used for comparing the relative performance of each strategy.

Table 10: Variables used for visualization

In this expression, $N_d = \delta_{H_D}(d)$ is the set of journals in discipline d , while $N_d(m) \subseteq N_d$ is the set of journals adopting strategy m in discipline d .

The function σ_d models the probability of a journal that gained δ positions in the ranking using strategy m being penalized as a fraction of a maximum probability equal to $r_{\text{detection}}$. The two terms in parentheses determine what fraction of the maximum probability will represent the risk of being penalized for the journal. If strategy m dominates discipline d , $\left(1 - \frac{|N_d(m)|}{|N_d|}\right)$ will be 0, and σ_d will be zero. On the other hand, $(1 - e^{-\Delta})$ will be zero if $\Delta = 0$ and approaches 1 as Δ grows.

Each journal using strategy m will have the sanction applied with a probability $\sigma_d(\Delta, m)$ for each discipline d in which it publishes. The exclusion of the journal from the ranking will last for t_{sanction} years, after which the journal is reinserted into the ranking. With the application of the sanction, the journal that occupied position p will now occupy position $|J| + 1$, thus having $\Delta = p - (|J| + 1)$. During the sanction period, the journal continues to publish, and the calculation of the JIF continues to be updated. However, the position of the journal remains always $|J| + 1$.

7.10 Update view

The visualization submodel is based on three elements:

1. Variable Plots: a list of aggregated variable plots that are interesting to observe in exploring the model.
2. Impact Factor Ranking: displays the real-time JIF ranking, which can also be exported for external analysis.
3. Network Visualization: graphs can be visualized at the level of articles, journals, or disciplines.

The aggregated variable plots to be considered in the model are detailed in Table 10, followed by the calculation from the state variables that must be performed to obtain them.

7.10.1 Update ranking

To update the Impact Factor ranking, the *output* state variable is formatted appropriately so that the data export submodel detailed in Section 7.11 can perform its function. At each simulation step, some aggregated data regarding each journal is collected to construct a ranking that closely resembles the SCImago ranking fields and format. Table 11 details the generated fields and how their calculation should be done from the state variables.

Algorithm 9 shows the pseudocode that generates the ranking.

Field Name	Type/Units	Meaning
Rank	Integer number	The position of the journal in the JIF ranking.
Title	Text	The title of the journal.
Total Docs. (1year)	Integer number	The total number of documents published by the journal in the current year, including all types of documents (articles, reviews, letters)
Total Docs. (2years)	Integer number	The total number of documents published by the journal in the previous 2 years, excluding the current year.
Citable Docs. (2years)	Integer number	The number of citable documents (articles and reviews) published by the journal in the previous 2 years, excluding the current year.
Total Cites (2years)	Integer number	The total number of citations received by any articles published by the journal in the previous 2 years, excluding the current year.
JIF (2years)	Real number	The impact factor calculated by dividing Total Cites (2years) by Citable Docs. (2years).
JIF Offset	Real number	The difference between the current and the previous JIF (2years) values.
Strategy	Text	The editorial strategy used by the journal.
Subject Areas	List	A list of all disciplines for which the journal publishes to, separated by semicolons.
Year	Integer number	The year of the current ranking.

Table 11: Description of the fields of the journal rankings generated when exporting data.

Algorithm 9 Update Journal Impact Factor ranking

```

procedure UPDATERANKING
  ranking  $\leftarrow$  ()
  for all  $j \in J$  ordered by incrising JIF do
    row  $\leftarrow$  GETFIELDS( $j$ )                                 $\triangleright$  Returns row with fields listed in Table 11 for  $j$ 
    ranking  $\leftarrow$  ranking + (row)                           $\triangleright$  Adds the new row to tuple ranking
  end for
  output  $\leftarrow$  ranking
end procedure

```

7.11 Export data

The journal ranking generated in real-time during the simulation execution can be exported as a CSV file for detail analysis. The ranking is exported for the entire simulated period, meaning that the CSV file will contain one ranking for each year simulated by the model. The fields available for export are detailed in Table 11. The heterogeneous temporal graph with disciplines, journals, documents, and their respective relationships can also be exported. In this case, however, the export is done only for the graph generated at the end of the simulated period, as the generated entities are cumulative, and each graph is a superset of the previous one. The exported graph can be loaded into specialized external tools such as Gephi or VOSviewer.

References

- [1] Volker Grimm, Uta Berger, Donald L. DeAngelis, J. Gary Polhill, Jarl Giske, and Steven F. Railsback. The ODD protocol: A review and first update. *Ecological Modelling*, 221(23):2760–2768, 2010. ISSN 0304-3800. doi: <https://doi.org/10.1016/j.ecolmodel.2010.08.019>. URL <https://www.sciencedirect.com/science/article/pii/S030438001000414X>.
- [2] Allen Wilhite, Eric A. Fong, and Seth Wilhite. The influence of editorial decisions and the academic network on self-citations and journal impact factors. *Research Policy*, 48(6):1513–1522, July 2019. ISSN 0048-7333. doi: 10.1016/J.RESPOL.2019.03.003. URL <https://acrobat.adobe.com/link/review?uri=urn:aaid:scds:US:9e5ecd8f-9368-3ce3-b46a-4471825eb1a2>. Publisher: North-Holland.
- [3] Christopher Watts and Nigel Gilbert. Does cumulative advantage affect collective learning in science? An agent-based simulation. *Scientometrics*, 89(1):437–463, October 2011. ISSN 1588-2861. doi: 10.1007/s11192-011-0432-8. URL <https://doi.org/10.1007/s11192-011-0432-8>.
- [4] David Magnus. Overthrowing the Tyranny of the Journal Impact Factor. *The American Journal of Bioethics*, 13(7):1–2, July 2013. ISSN 1526-5161. doi: 10.1080/15265161.2013.807709. URL <https://doi.org/10.1080/15265161.2013.807709>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/15265161.2013.807709>.
- [5] SCImago. SCImago Journal & Country Rank [Portal], 2021. URL <https://www.scimagojr.com/>.