Matthew Effects via Team Semantics

Sabine Frittella, Giuseppe M Greco, Michele Piazzai, Nachoem M Wijnberg, Fan M Yang

Matthew Effects via Team Semantics

SABINE FRITTELLA, INSA Centre Val de Loire, U. of Orléans, LIFO EA 4022, France
GIUSEPPE GRECO, Department of Values, Technology and Innovation, Delft U. of Technology, the Netherlands
MICHELE PIAZZAI, Department of Values, Technology and Innovation, Delft U. of Technology, the Netherlands
NACHOEM M. WIJNBERG, Amsterdam Business School, U. of Amsterdam, the Netherlands; Department of Pure and Applied Mathematics, U. of Johannesburg, South Africa
FAN YANG, Department of Values, Technology and Innovation, Delft U. of Technology, the Netherlands

Social scientists call *Matthew effect* the self-reinforcing mechanisms whereby initially small advantages accrued by individuals, e.g. in reputation, capital, or access to opportunities, beget further advantage and result in growing inequality. While there is extensive literature on the Matthew effect, the notion has not been explicitly defined. In this paper, we take a first step in this direction by providing a formalisation of Matthew effects within the framework of team semantics, and by introducing a logic for analysing the properties of the dependence relations involved in Matthew effects. We also show via an example how to use this logic to formalise a statistical analysis of a Matthew effect.

CCS Concepts: • **General and reference → General conference proceedings**; • **Information systems → Relational database model**; • **Theory of computation → Logic and databases**; *Data modeling*; • **Mathematics of computing →** *Regression analysis*; • **Applied computing →** *Decision analysis*;

Additional Key Words and Phrases: Matthew effect; Team Semantics; Dependence Logic.

ACM Reference format:
Sabine Frittella, Giuseppe Greco, Michele Piazzai, Nachoem M. Wijnberg, and Fan Yang. 2017. Matthew Effects via Team Semantics. 1, 1, Article 1 (April 2017), 13 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

We provide a logical formalisation of the social phenomenon termed *Matthew effect* using team semantics. First introduced by Merton [12], the term Matthew effect refers to the self-reinforcing process whereby reputationally rich academics tend to get richer over time. In Merton's words, it corresponds to "the accruing of large increments of peer recognition to scientists of great repute for particular contributions in contrast to the minimising or withholding of such recognition for scientists who have not yet made their mark" [13]. Outside the realm of academia, the Matthew effect has been invoked to explain positive feedbacks in e.g. the evaluation of athletes [10] and the conferral of public subsidies to firms [1]. Because of its ubiquity in social life, it has been recognized as a powerful engine of social, economic, and cultural inequality [14].

The role of Matthew effects is particularly evident in markets, where the uncertainty buyers face about the quality of products facilitates the consolidation of status hierarchies among sellers. This is because buyers consider the status of sellers as a good proxy for the quality of their products, so that higher status leads to greater visibility and access to resources, which help sellers achieve even greater status. For this reason, an extensive literature on the Matthew effect exists in sociology, economics, and management science. However, the Matthew effect is not precisely defined in this literature. As a result, researchers are hardly able to compare and integrate theoretical models and empirical findings. This motivates our present attempt to formalise the Matthew effect in logic.

The logical framework we propose for our formalisation is the framework of team semantics. Introduced originally by Hodges [8, 9] and later advanced by Väänänen [15], team semantics is a novel and effective logical tool for analysing the notion of *dependency* that is fundamental to the social and natural sciences. These dependencies usually manifest themselves only in the presence of multiple observations. Team semantics thus evalutes formulas under sets

DOI: 10.1145/nnnnnnn.nnnnnnn

of assignments (called *teams*), instead of single assignments as in the standard Tarskian semantics. Teams can be easily conceived as sets of rows in tables or data sets. The flexible and multidisciplinary interpretations of teams engendered a rapid development of logics based on team semantics in recent years. Notable such logics are *dependence logic* [15], *independence logic* [5] and *inclusion logic* [3, 4], which focus on characterising *functional dependence*, *independence* and *inclusion relations* among variables, respectively. In this paper we focus on the dependence relations relevant to Matthew effects. To the best of our knowledge, these have not been formalised in the literature. We take a first step in this direction by introducing a new logic based on team semantics, called *logic of Matthew effects* (ML), and analysing basic properties of these dependence relations within the framework of ML.

Structure of the paper. In Section 2, we define the syntax and semantics of our logic ML. In Section 3, we formalise four distinct types of Matthew effects and study their properties. In Section 4, we use ML to formalise an existing study on the Matthew effect in science [2].

2 LOGIC OF MATTHEW EFFECTS

We first present the general framework of the logic ML, including team semantics, mathematical definitions, and notation. We provide a brief presentation of time series regression, a common tool to analyse trends in dynamic data. For more details on this topic, the reader can refer to [6]. In Section 2.2, we introduce the basic dependence relations we use as building blocks to define Matthew effects. In Section 2.3, we present the syntax and semantics of ML.

2.1 General framework

We introduce the general framework to talk about data sets and regressions over data sets. In Appendix A.1, we present a toy example of a data set (Table 1) and a Matthew effect to exemplify the team semantics. We refer to this example throughout this subsection. In this paper, we assume that the empirical context where the Matthew effect occurs is described by a first-order model M of some signature $\mathscr{L}$ that will be specified in the sequel. The model M is assumed to have a three-sorted domain: *data sort*, *regression sort*, and *duration sort*.

Three-sorted domain. Elements of the *data sort* are all possible values in the data sets, e.g. real numbers or names. We include the value undefined to the possible values for variables of type data sort. We use $w, x, y, z, \dots$ (possibly with subscripts) to stand for variables of this sort. Data sets can be presented as tables (e.g. Table 1), and the variables of data sort can be understood as attributes in such tables. We assume that the time attribute is present in every data set, and we reserve the letter t for the time variable. In this setting, an observation in a data set or a row of a table corresponds to an assignment s that assigns to each variable x of data sort a value a of data sort in the domain of the underlying model M. Formally, a data set D consisting of several observations is a set of assignments, which we also call *team*. In this paper, the terms *data set* and *team* are used interchangeably. As we explained in the introduction, the logic we introduce for Matthew effects adopts team semantics, where formulas are evaluated on teams.

Elements of the *regression sort* are names of regressions, and we use $\mathsf{r}, \mathsf{r}', \dots$ to stand for variables of this sort. For a given model M, we call *granularity* the minimal time interval δ, between two observations. For instance, in Table 1 the granularity is $\delta = 1\ \textit{year}$. Different regressions can be performed on any data set using different variables. Consecutive observations for a single variable can be aggregated over time intervals that correspond to particular multiples of the granularity δ. The intervals of time (i.e., the natural numbers that correspond to the multiples of the minimal time interval) form the domain of the *duration sort*. We use $\ell, \ell', \dots$ to stand for variables of this sort.

Regression functions. Each regression generates one *regression function* for each dependent variable y under consideration. A regression function can be represented informally as:

$$y_{(\mathrm{t})} = \sum_{\{i_1,\dots,i_k\}\in\mathcal{X}} \beta_{i_1\dots i_k}(x_{i_1})_{(\mathrm{t}-\mathsf{l}\delta)}\cdots(x_{i_k})_{(\mathrm{t}-\mathsf{l}\delta)} + \epsilon, \tag{1}$$

where $\{x_1, \dots, x_n\}$ is a set of independent variables, $z_{(\mathrm{t})}$ is the value of variable z at time t, $\mathcal{X}$ is a non-empty downward closed subset of $\mathscr{P}(\{1, \dots, n\})$ (i.e., $B \subseteq A \in \mathcal{X}$ implies $B \in \mathcal{X}$), the positive real number $\mathsf{l} \cdot \delta$ is the time interval of the regression, l is a natural number, ϵ is an error term, and each coefficient $\beta_{i_1\dots i_k}$ is a real number. Most studies in the social sciences, including those on Matthew effects, use regressions of degree two or less, for instance:

$$y_{(\mathrm{t})} = \alpha + \beta_1(x_1)_{\mathrm{t}-\mathsf{l}\delta} + \beta_2(x_2)_{\mathrm{t}-\mathsf{l}\delta} + \beta_3(x_3)_{\mathrm{t}-\mathsf{l}\delta} + \beta_4(x_1)_{\mathrm{t}-\mathsf{l}\delta}(x_2)_{\mathrm{t}-\mathsf{l}\delta} + \beta_5(x_2)_{\mathrm{t}-\mathsf{l}\delta}(x_3)_{\mathrm{t}-\mathsf{l}\delta} + \epsilon. \tag{2}$$

In a data set D, if s and s' are two observations that respectively contain the data for time t and t−lδ (i.e., $s(t) = s'(t) + \mathsf{l} \cdot \delta$), then the above regression equation (2) can be formally represented as:

$$s(y) = \alpha + \beta_1 s'(x_1) + \beta_2 s'(x_2) + \beta_3 s'(x_3) + \beta_4 s'(x_1)s'(x_2) + \beta_5 s'(x_2)s'(x_3) + \epsilon.$$

If x_1 and x_2 are the focal variables, we will sometimes abbreviate the expression above as:

$$s(y) = \beta_1 s'(x_1) + \beta_2 s'(x_2) + \beta_4 s'(x_1)s'(x_2) + q(s'(x_1), s'(x_2), s'(x_3)).$$

Analysed data set. We call a data set D associated with regression $\mathsf{R}_1, \ldots, \mathsf{R}_\mathsf{k}$ with durations of time $\mathsf{l}_1, \ldots, \mathsf{l}_\mathsf{k}$, respectively, an *analysed data set*, denoted $(D, \mathsf{R}_1, \mathsf{l}_1, \ldots, \mathsf{R}_\mathsf{k}, \mathsf{l}_\mathsf{k})$. Formally, we view this analysed data set as a data set extended from D by adding $2k$ columns with attributes $\mathsf{r}_1, \ldots, \mathsf{r}_\mathsf{k}, \ell_1, \ldots, \ell_k$, where each column r_i has a constant value R_i, and each column ℓ_i has a constant value l_i. For simplicity, we denote this analysed data set also by D, and write $D(\mathsf{r}_\mathsf{i})$ for the unique value of the attribute r_i in D, and $D(\ell_i)$ for the unique value of the attribute ℓ_i. Technically, we may view a name R of a regression analysis as a function mapping each dependent variable y under consideration to a polynomial $\mathsf{R}(y)$ of the form (1).

2.2 Basic dependence relations.

We define basic dependence relations and atomic formulas that we later use to define Matthew effects.

Definition 2.1 (Basic dependence relations and their team semantics). Let $x_1, \ldots, x_n, y$ be data sort variables, r be a regression sort variable, and ℓ be a duration sort variable.

- The atomic formula $x_1, \ldots, x_n \nearrow^{\mathsf{r}}_{\ell} y$ characterises the notion of *y being **positively ℓ-dependent on** $x_1, \ldots, x_n$ with respect to the regression* r. We say that the defined relation is true in a data set D with an underlying model M, denoted $M \models_\mathsf{D} x_1, \ldots, x_n \nearrow^{\mathsf{r}}_{\ell} y$, iff for all $s, s' \in D$,

$$s(t) = s'(t) + D(\ell) \cdot \delta^M \Longrightarrow s(y) = \beta_1 s'(x_1) + \cdots + \beta_n s'(x_n) + q(s'(\vec{x}), s'(\vec{w})).$$

  where each β_i is significantly greater than 0 (see Appendix B.2), and $D(\mathsf{r})(y)$ is the polynomial represented above.
- The atomic formula $x_1, \ldots, x_n \searrow^{\mathsf{r}}_{\ell} y$ characterises the notion of *y being **negatively ℓ-dependent on** $x_1, \ldots, x_n$ with respect to* r, and we define $M \models_\mathsf{D} x_1, \ldots, x_n \searrow^{\mathsf{r}}_{\ell} y$ the same way as above except that we now require each β_i to be significantly smaller than 0.
- The atomic formula $x_1 \otimes \ldots \otimes x_n \nearrow^{\mathsf{r}}_{\ell} y$ characterises the notion of *y being **positively moderated ℓ-dependent on** $x_1, \ldots, x_n$ with respect to* r, and we define $M \models_\mathsf{D} x_1 \otimes \ldots \otimes x_n \nearrow^{\mathsf{r}}_{\ell} y$ iff for all $s, s' \in D$,

$$s(t) = s'(t) + D(\ell) \cdot \delta^M \Longrightarrow s(y) = \sum_{k=1}^{n} \sum_{1 \le i_1 < \cdots < i_k \le n} \beta_{i_1 \ldots i_k} s'(x_{i_1}) \cdots s'(x_{i_k}) + q(s'(\vec{x}), s'(\vec{w})),$$

  where $\beta_{i \ldots n}$ is significantly greater than 0, and $D(\mathsf{r})(y)$ is the polynomial represented as above.
- The atomic formula $x_1 \otimes \ldots \otimes x_n \searrow^{\mathsf{r}}_{\ell} y$ characterises the notion of *y being **negatively moderated ℓ-dependent on** $x_1, \ldots, x_n$ with respect to* r, and we define $M \models_\mathsf{D} x_1 \otimes \ldots \otimes x_n \searrow^{\mathsf{r}}_{\ell} y$ the same way as above except that we now require $\beta_{i \ldots n}$ to be significantly smaller than 0.

Definition 2.2 (Atomic formulas for positive and negative dependency). Let $\mathbb{X}$ denote the set of strings X of dependent variables of the form

$$x_{11} \otimes \cdots \otimes x_{1n_1}, \ \ldots, \ x_{k1} \otimes \cdots \otimes x_{kn_k}.$$

For every string $X \in \mathbb{X}$ such that $X = x_{11} \otimes \cdots \otimes x_{1n_1}, \ \ldots, \ x_{k1} \otimes \cdots \otimes x_{kn_k}$ and for every duration of time $\ell \cdot \delta$, we introduce the atomic formulas $X \nearrow^{\mathsf{r}}_{\ell} y$ and $X \searrow^{\mathsf{r}}_{\ell} y$, defined as follows:

- $M \models_D x_{11} \otimes \cdots \otimes x_{1n_1}, \ \ldots, \ x_{k1} \otimes \cdots \otimes x_{kn_k} \nearrow^{\mathsf{r}}_{\ell} y$ iff for all $s, s' \in D$,

$$s(t) = s'(t) + D(\ell) \cdot \delta^M \Longrightarrow s(y) = \sum_{i=1}^{k} \sum_{m=1}^{n_i} \sum_{1 \le j_1 < \cdots < j_m \le n_i} \beta_{i, j_1 \ldots j_m} s'(x_{ij_1}) \cdots s'(x_{ij_m}) + q(s'(\vec{x}), s'(\vec{w})), \tag{3}$$

  where each $\beta_{i,1,\ldots,n_i}$ is significantly greater than 0, and $D(\mathsf{r})(y)$ is the polynomial represented represented as above.
- $M \models_D x_{11} \otimes \cdots \otimes x_{1n_1}, \ \ldots, \ x_{k1} \otimes \cdots \otimes x_{kn_k} \searrow^{\mathsf{r}}_{\ell} y$ iff the same as the above holds except that each $\beta_{i,1,\ldots,n_i}$ is now required to be significantly smaller than 0.

2.3 The logic for Matthew effects

Here, we define the signature, the syntax, and the semantics of ML.

Definition 2.3 (The signature $\mathcal{L}$). The signature $\mathcal{L}$ of ML contains the constant functions 0 and δ, the unary functions $\{\mathsf{Effect}_{y,X} \mid y \in \mathsf{Var}, X \in \mathbb{X}\}$, the unary predicate Small, and the binary predicates $\gg$, $\ll$, and $\approx$.

The function symbols $\mathsf{Effect}_{y,X}$ are used to refer to certain combinations of the coefficients in the polynomials of the regression functions. For every variables x, y, every string X of dependent variables, $\mathsf{Effect}_{y,X}(\mathsf{r})$ is a regression sort argument, whose interpretation in the intended models M under a data set D is defined as:

- iff $X = x_{11} \otimes \cdots \otimes x_{1n_1}, \ldots, x_{k1} \otimes \cdots \otimes x_{kn_k}$, and $\beta_1 x_{11} \cdots x_{1n_1}$, ..., $\beta_k x_{k1} \cdots x_{kn_k}$ are terms in the polynomial $D(\mathsf{r})(y)$, then for every $s \in D$, $\mathsf{Effect}^M_{y,X}(s(\mathsf{r})) = \beta_1 + \cdots + \beta_k$. Otherwise, set $\mathsf{Effect}^M_{y,X}(s(\mathsf{r})) = \mathsf{undefined}$.

We equivalently denote $\mathsf{Effect}_{y,X}(\mathsf{r})$ by $\mathsf{Effect}(\mathsf{r}, y, X)$. The use of these terms becomes clearer in Section 4. The constant symbols 0 and δ are to be interpreted as the natural number 0 and the granularity of the data sets δ respectively. The predicate symbols Small, $\ll$, $\gg$ and $\approx$ are to be interpreted as "small," "significantly smaller than," "significantly greater than," and "equivalent to," respectively.

Definition 2.4 (Syntax). Let Var_0, Var_1 and Var_2 be respectively countable sets of variables of data sort, regression sort, and duration sort. The syntax of ML is defined as follows:

Terms of data sort	$\alpha ::= x \mid 0 \mid \delta \mid \mathsf{Effect}_{y,X}(\mathsf{r})$
Terms of duration sort	$\beta ::= \ell$
Terms of regression sort	$\gamma ::= \mathsf{r}$
Formulas	$\phi ::= X \nearrow^{\mathsf{r}}_{\ell} y \mid X \searrow^{\mathsf{r}}_{\ell} y \mid \mathsf{Small}(\alpha) \mid \alpha \ll \alpha \mid \alpha \gg \alpha \mid \alpha \approx \alpha \mid \phi \wedge \phi \mid \exists^1 x\, \phi \mid \exists^1 \ell\, \phi \mid \exists^1 \mathsf{r}\, \phi$

where $x \in \mathsf{Var}_0$, $\mathsf{r} \in \mathsf{Var}_1$ and $\ell \in \mathsf{Var}_2$.

Definition 2.5 (Free variables). For compound formulas, the sets of free variables of each sort are defined as usual. For atomic formulas, the sets of free variables of each sort are defined as follows:

- The set $\mathrm{Fv}_0(\phi)$ of free variables of data sort is defined as
  - for $X = x_{11} \otimes \cdots \otimes x_{1n_1}, \ldots, x_{k1} \otimes \cdots \otimes x_{kn_k}$,
  $$\mathrm{Fv}_0(X \nearrow^{\mathsf{r}}_{\ell} y) = \mathrm{Fv}_0(X \searrow^{\mathsf{r}}_{\ell} y) = \{x_{i,j} \mid 1 \le i \le k \text{ and } 1 \le j \le n_i\} \cup \{y, t\}^1,$$
  - $\mathrm{Fv}_0(\mathsf{Small}(\alpha)) = \mathrm{Fv}_0(\alpha)$,
  - $\mathrm{Fv}_0(\alpha \ll \beta) = \mathrm{Fv}_0(\alpha \gg \beta) = \mathrm{Fv}_0(\alpha \approx \beta) = \mathrm{Fv}_0(\alpha) \cup \mathrm{Fv}_0(\beta)$, and
  - the set $\mathrm{Fv}_0(\alpha)$ is defined inductively as $\mathrm{Fv}_0(x) = \{x\}$ and $\mathrm{Fv}_0(0) = \mathrm{Fv}_0(\delta) = \mathrm{Fv}_0(\mathsf{Effect}_{y,X}(\mathsf{r})) = \varnothing$.
- The set $\mathrm{Fv}_1(\phi)$ of free variables of regression sort is defined as
  - $\mathrm{Fv}_1(X \nearrow^{\mathsf{r}}_{\ell} y) = \mathrm{Fv}_1(X \searrow^{\mathsf{r}}_{\ell} y) = \{\mathsf{r}\}$,
  - $\mathrm{Fv}_1(\mathsf{Small}(\alpha)) = \mathrm{Fv}_1(\alpha)$ and $\mathrm{Fv}_1(\alpha \ll \beta) = \mathrm{Fv}_1(\alpha \gg \beta) = \mathrm{Fv}_1(\alpha \approx \beta) = \mathrm{Fv}_1(\alpha) \cup \mathrm{Fv}_1(\beta)$, and
  - the set $\mathrm{Fv}_1(\alpha)$ is defined inductively as $\mathrm{Fv}_1(x) = \mathrm{Fv}_1(0) = \mathrm{Fv}_1(\delta) = \varnothing$ and $\mathrm{Fv}_1(\mathsf{Effect}_{y,X}(\mathsf{r})) = \{\mathsf{r}\}$.
- The set $\mathrm{Fv}_2(\phi)$ of free variables of duration sort is defined as
  - $\mathrm{Fv}_2(X \nearrow^{\mathsf{r}}_{\ell} y) = \mathrm{Fv}_2(X \searrow^{\mathsf{r}}_{\ell} y) = \{\ell\}$ and
  - $\mathrm{Fv}_2(\phi) = \varnothing$ for any other atomic formula.

Formulas with sets $\mathsf{Var}_0, \mathsf{Var}_1, \mathsf{Var}_2$ of free variables of data sort, regression sort, and duration sort are evaluated on a model M with respect to *teams* over $V_0 \cup V_1 \cup V_2$, i.e., *sets* D of assignments $s : V_0 \cup V_1 \cup V_2 \to M$.

Definition 2.6 (Semantics). We define inductively the satisfaction relation $M \models_{\mathrm{D}} \phi$ as follows:

- See Theorem 2.2 for the team semantics of the atomic formulas $X \nearrow^{\mathsf{r}}_{k\ell} y$ and $X \searrow^{\mathsf{r}}_{k\ell} y$.
- For the other atomic formula θ, $M \models_D \theta$ iff $M \models_s \theta$ in the usual sense for all $s \in D$.
- $M \models_D \phi \wedge \psi$ iff $M \models_D \phi$ and $M \models_D \psi$.
- $M \models_D \exists^1 x\, \phi$ iff $M \models_{D(a/x)} \phi$ for some element $a \in M$ of data sort, where $D(a/x) = \{s(a/x) \mid s \in D\}$.
- $M \models_D \exists^1 \mathsf{r}\, \phi$ and $M \models_D \exists^1 \ell\, \phi$ are defined as above respecting the sorts of the variables.

[1] Notice that we assume the variable t to be always present in these atoms, although for simplicity we do not explicitly write the variable in the formulas. Cf. the team semantics given in (3).

For any set $\Gamma \cup \{\phi\}$ of formulas, we write $\Gamma \vDash \phi$ if for all models M and teams D, $M \vDash_D \gamma$ for all $\gamma \in \Gamma$ implies $M \vDash_D \phi$. We write $\phi \vDash \psi$ for $\{\phi\} \vDash \psi$.

3 FORMALISING MATTHEW EFFECTS

In this section, we formalise four distinct types of Matthew effects in ML and we investigate the properties of the dependence relations and Matthew effects.

Definition 3.1 (Matthew effects). We define the following notions:

- *y being subject to a **positive direct ℓ-Matthew effect** with respect to* r (see also Table 2(b)):

$$\mathsf{DME}^{\mathsf{r}}_{\ell} y ::= y \nearrow^{\mathsf{r}}_{\ell} y\,.$$

- *y being subject to a **positive x-mediated ℓ-Matthew effect** with respect to* r (see also Table 2(c)):

$$\mathsf{MME}^{\mathsf{r}}_{\ell} y(x) ::= x \nearrow^{\mathsf{r}}_{\ell} y \wedge y \nearrow^{\mathsf{r}}_{\ell} x\,.$$

- *y being subject to a **positive x-complete ℓ-Matthew effect** with respect to* r (see also Table 2(d)):

$$\mathsf{CME}^{\mathsf{r}}_{\ell} y(x) ::= \mathsf{MME}^{\mathsf{r}}_{\ell} y(x) \wedge \mathsf{DME}^{\mathsf{r}}_{\ell} y.$$

- *x and y being subjects to a **positive complete ℓ-Matthew effect** with respect to* r (see also Table 2(e)):

$$\mathsf{CME}^{\mathsf{r}}_{\ell}(x, y) ::= \mathsf{MME}^{\mathsf{r}}_{\ell} y(x) \wedge \mathsf{DME}^{\mathsf{r}}_{\ell} x \wedge \mathsf{DME}^{\mathsf{r}}_{\ell} y.$$

Properties. It is not hard to verify that the dependence relation $X \nearrow^{\mathsf{r}}_{\ell} y$ satisfies the following properties:

- (Reflexivity) $\vDash x \nearrow^{\mathsf{r}}_{0} x$
- (Enhancing) $x \nearrow^{\mathsf{r}}_{\ell} x \vDash \exists^1 \mathsf{r}\ x \nearrow^{\mathsf{r}}_{k\ell} x$
- (Commutativity) $X_1, \cdots, X_n \nearrow^{\mathsf{r}}_{\ell} y \vDash X_{i_1}, \cdots, X_{i_n} \nearrow^{\mathsf{r}}_{\ell} y$ and $W, x_1 \otimes \cdots \otimes x_n, Z \nearrow^{\mathsf{r}}_{\ell} y \vDash W, x_{i_1} \otimes \cdots \otimes x_{i_n}, Z \nearrow^{\mathsf{r}}_{\ell} y$, where $i_1, \ldots, i_n$ is any permutation of $1, \ldots, n$
- (Duplication) $X_1, \ldots, X_n \nearrow^{\mathsf{r}}_{\ell} y \vDash X_i, X_1, \ldots, X_n \nearrow^{\mathsf{r}}_{\ell} y$, where $X_i \in \{X_1, \ldots, X_n\}$
- (Projection) $X_1, \ldots, X_n \nearrow^{\mathsf{r}}_{\ell} y \vDash X_{i_1}, \ldots, X_{i_k} \nearrow^{\mathsf{r}}_{\ell} y$, where $X_{i_1}, \ldots, X_{i_k}$ is any subsequence of $X_1, \ldots, X_n$
- (Regrouping) $(X \nearrow^{\mathsf{r}}_{\ell} y), (Z \nearrow^{\mathsf{r}}_{\ell} y) \vDash X, Z \nearrow^{\mathsf{r}}_{\ell} y$
- (Transitivity) $(X \nearrow^{\mathsf{r}}_{\ell} y),\ (y \nearrow^{\mathsf{r}}_{k\ell} z) \vDash \exists^1 \mathsf{r}\ X \nearrow^{\mathsf{r}}_{(k+1)\ell} z$

As a consequence of the transitivity of the dependence relation, mediated and direct Matthew effects satisfy the following properties:

- (Transitivity) $\mathsf{MME}^{\mathsf{r}}_{\ell} x(y), \mathsf{MME}^{\mathsf{r}}_{\ell} y(z) \vDash \exists^1 \mathsf{r}\ \mathsf{MME}^{\mathsf{r}}_{2\ell} x(z)$
- (Scaling) $\mathsf{MME}^{\mathsf{r}}_{\ell} y(x) \vDash \exists^1 \mathsf{r}_0\ \mathsf{DME}^{\mathsf{r}_0}_{2\ell} x \wedge \exists^1 \mathsf{r}_1\ \mathsf{DME}^{\mathsf{r}_1}_{2\ell} y$

Moreover, a direct Matthew effect of a variable x is clearly a mediated Matthew effect where x itself is the mediator, namely, $\mathsf{DME}^{\mathsf{r}}_{\ell} y \vDash \mathsf{MME}^{\mathsf{r}}_{\ell} y(y)$, and a mediated Matthew effect is *reciprocal* for the two variables involved, namely $\mathsf{MME}^{\mathsf{r}}_{\ell} y(x) \vDash \mathsf{MME}^{\mathsf{r}}_{\ell} x(y)$.

4 CASE STUDY

In Section 4.1, we informally present the results of Azoulay, Stuart, and Wang (henceforth: ASW) reported in [2] about the Matthew effect in science. In Section 4.2, we formalise their analysis.

4.1 Informal presentation

ASW propose an empirical test for the following proposition: scientists of higher status will have even higher status in the future. According to the definitions presented in Section 2, this is equivalent to saying that a direct Matthew effect exists with regard to a scientist's status. The empirical test revolves around the conferral to medical scientists of the prestigious title of Howard Hughes Medical Investigator (HHMI). ASW examine how the yearly number of citations for an article published by a HHMI-appointed scientist before the appointment changes as a result of the appointment. ASW assume that both the HHMI appointment and an article's yearly number of citations reflect the scientist's status.

The hypothesis that a direct Matthew effect exists with regard to a scientist's status is accepted if the articles published by HHMI appointees receive more citations after the appointment, compared to articles of similar quality published by non-appointees. The results of the analysis show that:

- The citation boost is small and it affects only the articles published up to 1 year before the HHMI appointment. Older articles do not witness a change in citations received as a result of the appointment.
- The citation boost is larger for articles published in journals with low impact factor, articles that use more novel keywords, and articles that cite a greater number of studies from other fields (i.e., that are *recombinant*). ASW argue that this is because the quality of these articles is more difficult to assess; therefore, the HHMI appointment acts as a signal of quality and more strongly affects the yearly citations these articles receive.
- The citation boost is larger for articles published by scientists who have a smaller total number of citations attached to their name or who are younger at the time of the HHMI appointment.

On the basis of these results, ASW conclude that there is a Matthew effect with regard to scientists' status, but the extent to which this is observable depends on the age of the articles published by scientists and on how easily the quality of these articles can be assessed. In addition, they conclude that the Matthew effect more strongly affects scientists who have lower status at the time they are appointed.

4.2 Formalisation

ASW perform a complex empirical test involving multiple variables and regression models. These are presented in detail in Appendix B. In this subsection, we first formalise ASW's statements in isolation, then we analyse the reasoning they use to draw their conclusions.

Statements. Each statement is presented in natural language (*text in italic*) and then formalised using ML.

ASW aim at proving that *there is a direct Matthew effect with regard to a scientist's status*. This can be formalised by the formula:

$$\exists^1 \ell \, \exists^1 \mathrm{r}_a \quad \mathsf{DME}^{\mathrm{r}_a}_{\ell} \mathsf{Status}. \tag{4}$$

ASW observe (regression r_1) that ACitAF *positively depends on* HHMI*, but the effect is small* [2, page 21], which can be formalised by the formula:

$$\mathsf{HHMI} \nearrow^{\mathrm{r}_1}_{year} \mathsf{ACitAF} \quad \wedge \quad \mathsf{Small}(\mathrm{r}_1, \mathsf{ACitAF}, \{\mathsf{HHMI}\}). \tag{5}$$

In addition, they observe (regressions r_2, r_3, and r_4) that *this positive dependency only affects the articles published up to 1 year before the HHMI appointment*:

$$\mathsf{HHMI} \nearrow^{\mathrm{r}_2}_{year} \mathsf{ACitAF} \tag{6}$$

$$\wedge \quad \mathsf{Effect}(\mathrm{r}_2, \mathsf{ACitAF}, \{\mathsf{HHMI}\}) \gg \mathsf{Effect}(\mathrm{r}_1, \mathsf{ACitAF}, \{\mathsf{HHMI}\}) \tag{7}$$

$$\wedge \quad \mathsf{Effect}(\mathrm{r}_3, \mathsf{ACitAF}, \{\mathsf{HHMI}\}) \approx 0 \quad \wedge \quad \mathsf{Effect}(\mathrm{r}_4, \mathsf{ACitAF}, \{\mathsf{HHMI}\}) \approx 0. \tag{8}$$

Recall that the regressions r_2, r_3 and r_4 are respectively based on articles published up to 1 year, 2 years, and 3 to 10 years before the HHMI appointment (see Table 4).

ASW also observe (regression r_5) that *there is a stronger increase in citations after the HHMI appointment if the article is published in a journal with low impact factor*:

$$\mathsf{HHMI} \otimes \mathsf{LIF} \nearrow^{\mathrm{r}_5}_{year} \mathsf{ACitAF} \tag{9}$$

Furthermore, they observe that *there is a stronger increase in citations if the article is novel* (regression r_6):

$$\mathsf{HHMI} \otimes \mathsf{Novel} \nearrow^{\mathrm{r}_6}_{year} \mathsf{ACitAF}, \tag{10}$$

or if it is recombinant (regression r_7):

$$\mathsf{HHMI} \otimes \mathsf{Recombinant} \nearrow^{\mathrm{r}_7}_{year} \mathsf{ACitAF}\,. \tag{11}$$

ASW assume that *the quality of articles is more uncertain if they are published in journals with low impact factor, if they are novel, or if they are recombinant*:

$$\exists^1 \ell \, \exists^1 \mathrm{r}_b \quad \left(\{\mathsf{LIF}, \mathsf{Novel}, \mathsf{Recombinant}\} \nearrow^{\mathrm{r}_b}_{\ell} \mathsf{UArtQ}\right). \tag{12}$$

The assumption (12) and the observations (9), (10) and (11) suggest that *the positive dependency of* ACitAF *on* HHMI *more strongly affects articles of uncertain quality*:

$$\exists^1 \ell \, \exists^1 \mathrm{r}_c \quad \left(\mathsf{HHMI} \nearrow^{\mathrm{r}_c}_{\ell} \mathsf{ACitAF} \quad \wedge \quad \mathsf{HHMI} \otimes \mathsf{UArtQ} \nearrow^{\mathrm{r}_c}_{\ell} \mathsf{ACitAF}\right). \tag{13}$$

Moreover, ASW observe that *there is a stronger increase in the number of citations after the HHMI appointment if the article is published by a scientist who is less cited at the time of the appointment* (regression r_8):

$$\mathsf{HHMI} \otimes \mathsf{Hnotwellcited} \nearrow^{\mathsf{r}_8}_{year} \mathsf{ACitAF}\,. \tag{14}$$

or who is younger at the time of the appointment (regression r_9):

$$\mathsf{HHMI} \otimes \mathsf{Hyoung} \nearrow^{\mathsf{r}_9}_{year} \mathsf{ACitAF}\,. \tag{15}$$

The assumption that *the variables* Hnotwellcited *and* Hyoung *have a negative effect on* Status,

$$\exists^1 \ell\, \exists^1 \mathsf{r}_d \quad \big(\{\mathsf{Hnotwellcited}, \mathsf{Hyoung}\} \searrow^{\mathsf{r}_d}_{\ell} \mathsf{Status}\big)\,, \tag{16}$$

leads to the conclusion that *the positive dependency of* ACitAF *on* HHMI *more strongly affects scientists who have lower status at the time of the HHMI appointment than it affects scientists who have higher status*:

$$\exists^1 \ell\, \exists^1 \mathsf{r}_e \quad \big(\mathsf{HHMI} \nearrow^{\mathsf{r}_e}_{\ell} \mathsf{ACitAF} \quad \wedge \quad \mathsf{HHMI} \otimes \mathsf{Status} \searrow^{\mathsf{r}_e}_{\ell} \mathsf{ACitAF}\big)\,. \tag{17}$$

Reasoning. The observations (5), (6), (7), (8) (9), (10), (11), (14) and (15) and the assumptions (12) and (16) are considered as axioms. Here we list the deduction steps ASW use to deduce (4), (13) and (17). To deduce (13), ASW use the following reasoning: *if* ACitAF *is subject to a positively moderated year-dependency on* HHMI ⊗ LIF *(resp.* HHMI ⊗ Novel *or* HHMI ⊗ Recombinant*) witnessed by the regression* r*, and if* UArtQ *is subject to a positive* LIF *(resp.* Novel *or* Recombinant*) dependency, then there is an hypothetical regression* r′ *that witnesses the fact that* ACitAF *is subject to a positively moderated* ℓ'*-dependency on* HHMI ⊗ UArtQ. This reasoning is captured by the following deduction step:

$$\frac{h \otimes n \nearrow^{\mathsf{r}}_{\ell} c \qquad \exists^1 \ell\, \exists^1 \mathsf{r}\ n \nearrow^{\mathsf{r}}_{\ell} u}{\exists^1 \ell'\, \exists^1 \mathsf{r}'\ h \otimes u \nearrow^{\mathsf{r}'}_{\ell'} c}\,.$$

To deduce (17), ASW use the following reasoning: *if* ACitAF *is subject to a positively moderated year-dependency on* HHMI ⊗ Hnotwellcited *(resp.* HHMI ⊗ Hyoung*) witnessed by the regression* r_8 *(resp.* r_9*), and if* Hnotwellcited *(resp.* Hyoung*) has a negative effect on* Status*, then there is an hypothetical regression* r′ *that witnesses the fact that* ACitAF *is subject to a positively moderated* ℓ'*-dependency on* HHMI ⊗ Status. This reasoning is captured by the following deduction step:

$$\frac{h \otimes n \nearrow^{\mathsf{r}}_{\ell} c \qquad \exists^1 \ell\, \exists^1 \mathsf{r}\ n \searrow^{\mathsf{r}}_{\ell} s}{\exists^1 \ell'\, \exists^1 \mathsf{r}'\ h \otimes s \searrow^{\ell'}_{\mathsf{r}'} c}\,.$$

To deduce (4), one first need to express the assumption that HHMI *is positively dependent on* Status*:*

$$\exists^1 \ell\, \exists^1 \mathsf{r}_f \quad \big(\mathsf{Status} \nearrow^{\mathsf{r}_f}_{\ell} \mathsf{HHMI}\big)\,. \tag{18}$$

Based on this, ASW use the following reasoning: *if* HHMI *is positively dependent on* Status*, if* ACitAF *is positively dependent on* HHMI *as witnessed by the regression* r_1*, then* Status *is positively dependent on* Status*, which means that* Status *is subject to a direct Matthew effect.* This reasoning is captured by the following deduction steps:

$$\frac{\exists^1 \ell\, \exists^1 \mathsf{r}\ s \nearrow^{\mathsf{r}}_{\ell} h \qquad h \nearrow^{\mathsf{r}_1}_{year} c}{\exists^1 \ell\, \exists^1 \mathsf{r}\ s \nearrow^{\mathsf{r}}_{\ell} c} \quad \text{and} \quad \frac{\exists^1 \ell\, \exists^1 \mathsf{r}\ s \nearrow^{\mathsf{r}}_{\ell} c \qquad \exists^1 \ell\, \exists^1 \mathsf{r}\ c \nearrow^{\mathsf{r}}_{\ell} s}{\exists^1 \ell\, \exists^1 \mathsf{r}\ \mathsf{DME}^{\mathsf{r}}_{\ell} s.}$$

5 CONCLUSION AND FURTHER RESEARCH

Conclusion. While there is a great deal of literature in the social sciences invoking the Matthew effect to explain important phenomena, from career dynamics to economic inequality, the concept of Matthew effect has never been properly formalised. This makes it difficult to compare and synthesise the results of different studies. This paper offers a first formalisation of the Matthew effect via a logic based on team semantics.

An original contribution of this paper is that our formalisation allows for a clear distinction between different types of Matthew effects: direct, mediated, and complete. This shows just how complicated self-reinforcing phenomena can be, because an observed Matthew effect can actually result from the interplay of direct and mediated Matthew effects. In addition, our formalisation serves a number of interrelated purposes: first, as shown in our case study, it can be used to better understand the import of empirical research and to make explicit the assumptions needed to support the researchers' conclusions; second, it can be used to compare and relate the results of different studies to one another, and thereby develop new theory on a firmer foundation; third, it allows empirical scientists to ask new research questions and formulate more precise hypotheses.

This study is only a first step in exploring logical formalisms that address the intricate phenomena concerning the Matthew effect. Further work will expand the logical analysis in the following directions:

Studying properties of ML. The logic ML, we introduced is defined on the basis of team semantics. A team or a data set is essentially a relation of the model, which is a second-order object. As a consequence, logics based on team semantics are usually second-order in expressive power. Indeed, the two major team-based logics, dependence logic and independence logic, are expressibly equivalent to existential second-order logic [3, 15]. The atomic dependency notions formalised in ML are more involved than the functional dependency, independence, and other dependency notions studied so far in team-based logics. Yet the language of our ML is very simple, as the only complex formulas are the conjunctions and the existentially quantified statements with weak existential quantifiers of team semantics. One natural conjecture would be that this logic is strictly weaker in expressive power than existential second-order or even first-order logic.

A richer language with a good proof calculus. Although we demonstrated in the case study that the simple language of ML can already express interesting facts about Matthew effects, in future work we will introduce stronger logics by expanding the language to include disjunction, negation, implication, and strong quantifiers. We also want to introduce proof calculi for these extensions or their sufficiently strong fragments.

Reflecting the complexity of empirical tests. Our formalisation suggests that empirical findings about Matthew effects are contingent on particular statistical analyses, performed on particular data, where particular variables are observed over particular time intervals. Choices related to research design can thus affect the empirical evidence researchers find about Matthew effects. For example, choosing to observe the variables yearly rather than monthly, weekly, or daily can determine whether a direct Matthew effect is found in the place of a mediated one. Moreover, the fact that some variables like quality and uncertainty about quality remain unobserved can conceal important dependencies; as a result, a Matthew effect may appear to be mediated by a certain (observed) variable whereas in fact it is mediated by another (unobserved) one, which depends or is dependent on the apparent mediator. In formalising Matthew effects, or indeed any dependency tested via statistical analysis, one must be able to express these details within the syntax of the logic.

REFERENCES

[1] C Antonelli and F Crespi. 2013. The 'Matthew effect' in R&D public subsidies: The Italian evidence. *Technological Forecasting and Social Change* 80, 8 (2013), 1523–1534.

[2] P Azoulay, T Stuart, and Y Wang. 2013. Matthew: Effect or fable? *Management Science* 60, 1 (2013), 92–109.

[3] Pietro Galliani. 2012. Inclusion and Exclusion in Team Semantics: On some logics of imperfect information. *Annals of Pure and Applied Logic* 163, 1 (January 2012), 68–84.

[4] Pietro Galliani and Lauri Hella. 2013. Inclusion Logic and Fixed Point Logic. In *Computer Science Logic 2013 (CSL 2013), CSL 2013, September 2-5, 2013, Torino, Italy.* 281–295. DOI: https://doi.org/10.4230/LIPIcs.CSL.2013.281

[5] Erich Grädel and Jouko Väänänen. 2013. Dependence and Independence. *Studia Logica* 101, 2 (April 2013), 399–410.

[6] J. Hamilton. 1994. *Time Series Analysis.* Princeton University Press.

[7] Miika Hannula. 2015. Axiomatizing first-order consequences in independence logic. *Annals of Pure and Applied Logic* 166, 1 (2015), 61–91.

[8] W. Hodges. 1997. Compositional Semantics for a Language of Imperfect Information. *Logic Journal of the IGPL* 5 (1997), 539–563.

[9] W. Hodges. 1997. Some Strange Quantifiers. In *Structures in Logic and Computer Science: A Selection of Essays in Honor of A. Ehrenfeucht*, J. Mycielski, G. Rozenberg, and A. Salomaa (Eds.). Lecture Notes in Computer Science, Vol. 1261. London: Springer, 51–65.

[10] J W Kim and B G King. 2014. Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science* 60, 11 (2014), 2619–2644.

[11] Juha Kontinen and Jouko Väänänen. 2013. Axiomatizing first-order consequences in dependence logic. *Annals of Pure and Applied Logic* 164, 11 (2013).

[12] R K Merton. 1968. The Matthew effect in science. *Science* 159, 3810 (1968), 56–63.

[13] R K Merton. 1988. The Matthew effect in science, II: Cumulative advantage and the symbolism of Intellectual Property. *Isis* 79, 4 (1988), 606–623.

[14] D Rigney. 2010. *The Matthew effect: How advantage begets further advantage.* Columbia University Press, New York City, NY.

[15] Jouko Väänänen. 2007. *Dependence Logic: A New Approach to Independence Friendly Logic.* Cambridge: Cambridge University Press.

[16] Fan Yang. 2016. Negation and Partial Axiomatizations of dependence and independence logic revisited. In *Proceedings of 23rd Workshop on Logic, Language, Information and Computation (LNCS 9803)*, Jouko Väänänen, Asa Hirvonen, and Ruy de Queiroz (Eds.). Springger-Verlag, 410–431.

A APPENDIX - MATTHEW EFFECTS

A.1 Toy example of a Matthew effect

Matthew effects are often detected by researchers while analyzing empirical data. In the statistical literature, these represent a form of autocorrelation [6]. Table 1 presents a hypothetical dataset that shows *prima facie* evidence of a Matthew effect. This data concerns the careers of visual artists: each row contains information about an artist during a given year. The first column includes the artist ID; the second column includes the number of artworks sold by the artist during the observation year; the third column includes the number of times the artist or her work were reviewed by the media during the observation year; the fourth column specifies the observation year. Assume that all artists started their career in 2010 and that they were equally productive during the study period. The data suggests that Artist A started accumulating reviews from the very beginning, and sales quickly followed. A similar pattern can be observed for Artist B, though with some delay. In the case of Artist C, however, this trend never began.

We may presume that both sales and reviews are subject to a Matthew effect because selling more artworks leads to greater odds of selling artwork in the future. Similarly, being reviewed increases the odds of future reviews. However, it is also possible—and indeed highly likely—that being reviewed increases the odds of future sales, and that selling artwork increase the odds of future reviews. The fact that these dependencies occur at the same time makes the individual effects difficult to isolate.

Table 1. A data set (or a team)

Artist	Sales	Reviews	Time	...
A	0	1	2010	...
A	1	2	2011	...
A	1	1	2012	...
A	2	4	2013	...
A	4	7	2014	...
A	7	9	2015	...
⋮	⋮	⋮	⋮	

Artist	Sales	Reviews	Time	...
B	0	0	2010	...
B	0	0	2011	...
B	0	1	2012	...
B	0	2	2013	...
B	2	5	2014	...
B	4	8	2015	...
⋮	⋮	⋮	⋮	

Artist	Sales	Reviews	Time	...
C	0	0	2010	...
C	1	1	2011	...
C	0	0	2012	...
C	0	1	2013	...
C	1	2	2014	...
C	0	1	2015	...
⋮	⋮	⋮	⋮	

A.2 Different Matthew effects

Table 2 contains a graphic representation of the definitions of the positive dependency introduced in Theorem 2.1 and of the Matthew effects defined in Section 3. To help understand the definitions of the different Matthew effects and Table 2, we present a simple example of positive dependency.

Example A.1. Consider a very simple dependence relation $x \nearrow^{\mathsf{r}}_{\ell} y$ with a linear regression function

$$y_{(\mathrm{t})} = \alpha + \beta x_{(\mathrm{t}-\mathrm{l}\delta)} + \gamma_1 w_{1_{(\mathrm{t}-\mathrm{l}\delta)}} + \cdots + \gamma_1 w_{1_{(\mathrm{t}-\mathrm{l}\delta)}} + \epsilon,$$

where β is significantly greater than 0. Since we also have

$$y_{(\mathrm{t}-\mathrm{l}\delta)} = \alpha + \beta x_{(\mathrm{t}-2\mathrm{l}\delta)} + \gamma_1 w_{1_{(\mathrm{t}-2\mathrm{l}\delta)}} + \cdots + \gamma_1 w_{1_{(\mathrm{t}-2\mathrm{l}\delta)}} + \epsilon,$$

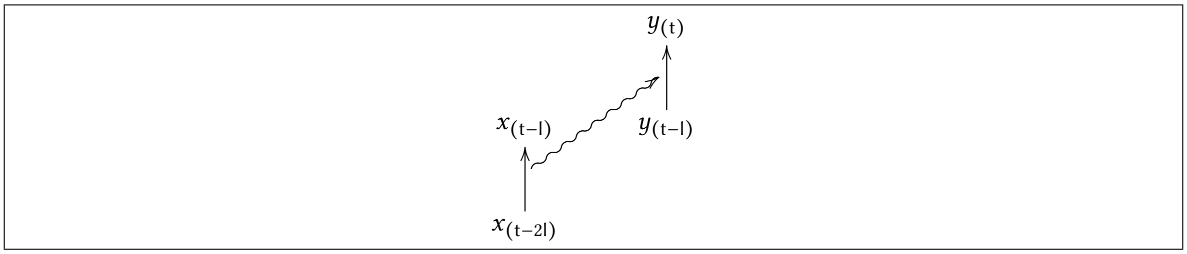
when all the independent variables except x are held constant, we obtain

$$y_{(\mathrm{t})} - y_{(\mathrm{t}-\mathrm{l}\delta)} = \beta\big(x_{(\mathrm{t}-\mathrm{l}\delta)} - x_{(\mathrm{t}-2\mathrm{l}\delta)}\big),$$

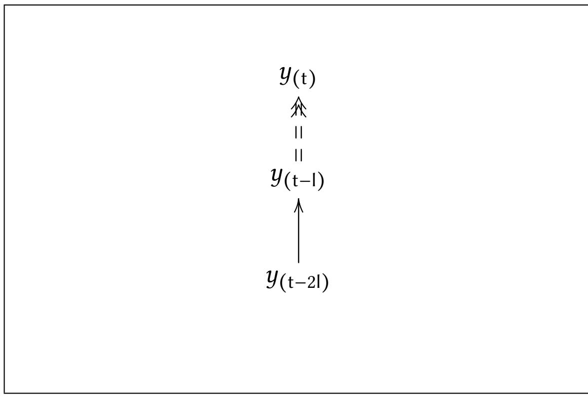
meaning that the value of y increases as the value of x increases (see also Table 2(a)).

Table 2. Types of Matthew effect

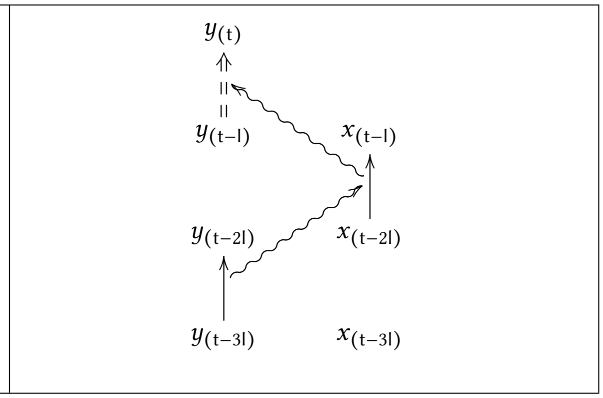
(a) Dependence relation



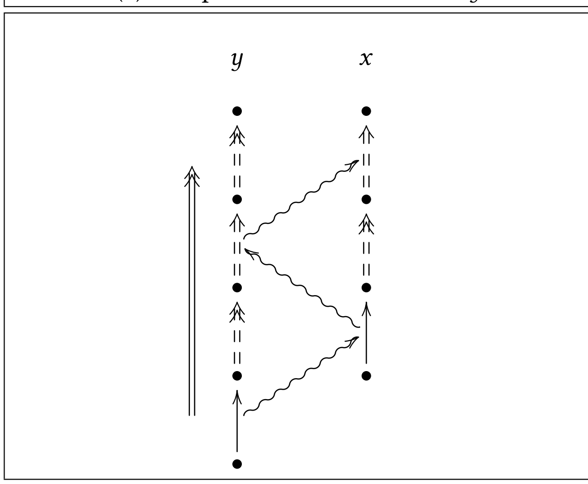
(b) Direct Matthew effect



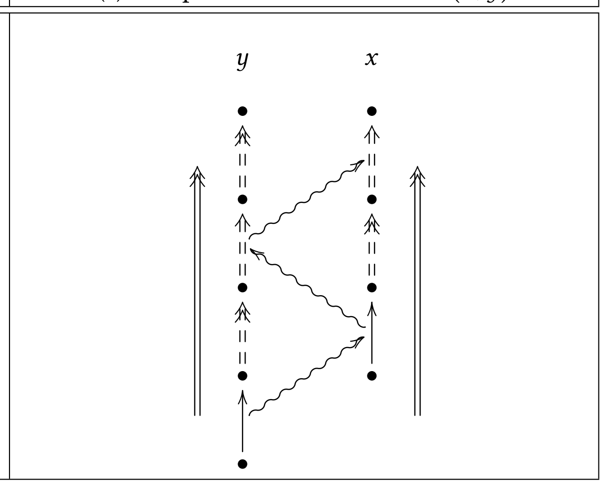
(c) Mediated Matthew effect



(d) Complete Matthew effect for y



(e) Complete Matthew effect for (x, y)



B APPENDIX - CASE STUDY

B.1 Variables

The key variables used by ASW to analyze the Matthew effect among scientists in [2] is reported in Table 3. Some of these variables measure relevant characteristics of a scientist, including Author, HHMI, Hdate, Hage, and Hcit. Other variables measure characteristics related to the article published by the scientist, including Article, ArtY, Journal, NbAut, Apos, ACitAF, ACitBF, IF, Novelty, and Recombination.

In [2], the variables Hage, Hcit, IF, Novelty, and Recombination are split along the median value observed in the data, giving rise to the following boolean variables: Hyoung, which is true if the scientist is younger than the median at the time of the HHMI appointment; Hwellcited, which is true if the scientist has a greater number of total citations than the median at the time of the HHMI appointment; HIF, which is true if the journal where the article is published has higher impact factor than the median; Novel, which is true if the keywords associated with the article are more novel than the median; and Recombinant, which is true if the proportion of out-of-field literature cited by the article is greater than the median.

In addition to these variables, which are measured and actually used in ASW's statistical analysis, the claims made in [2] about the Matthew effect involve a number of additional variables, which cannot be measured directly but are assumed to depend on some of the observed variables. These include Status, i.e., the status of the focal scientist, as well as ArtQ and UArtQ, which represent the quality of the scientist's article and the uncertainty about the quality of the scientist's article, respectively.

B.2 Regressions

A regression is performed via the application of an algorithm, i.e., an *estimator*, to the observed data. The algorithm yields a set of *coefficients*, which correspond to the β mentioned in Section 2. Each coefficient represents an *effect*, i.e., a change in the value of the dependent variable y that results from a one-unit increase in the value of the independent variable x. Each coefficient is associated with a *level of statistical significance*, i.e., a value between 0 and 1 that represents the probability of observing the estimated effect in the data. The lower this value, the greater the probability of observing the effect. A coefficient with a level of statistical significance below some predetermined threshold is said to be *significantly greater than* 0 if the coefficient is positive, and *significantly smaller than* 0 if the coefficient is negative. If the level of statistical significance is above the predetermined threshold, the coefficient is said to be *non-significant* or *equivalent to zero*. In [2], the chosen significance threshold is 0.05, as is conventional in the social sciences.

Table 4 reports the full list of the regressions performed by ASW in [2]. These are indexed by numbers 1–9. These regressions concern the observed variables listed in Table 3. In addition, Table 4 reports a list regressions that are not actually performed by ASW in [2], but their results are nonetheless relevant to ASW's analysis. We call these *hypothetical regressions*. These are not actually performed because they concern the unobserved variables listed in Table 3. However, they could be performed if it were possible to observe these variables. These hypothetical regressions are indexed by letters a–e. In every regression, the dependent variable is ACitAF.

Table 3. List of variables

		Variable	Meaning	Type	Time dependent
Author related	**Observed**	Author	Scientist ID	ℕ	
		HHMI	Author is appointed HHMI	0/1	
		Hdate	Date of the HHMI appointment	ℕ	
		Hage	Age of the author at HHMI appointment	ℕ	
		Hyoung	Author was young at HHMI appointment	0/1	
		Hold	Author was old at HHMI appointment	0/1	
		Hcit	Author's total citations at HHMI appointment	ℕ	
		Hwellcited	Author was well cited at HHMI appointment	0/1	
		Hnotwellcited	Author was not well cited at HHMI appointment	0/1	
	Unobserved	Status	Status of the author	ℝ	
Article related	**Observed**	Article	Article ID	ℕ	
		ArtY	Date of article publication	ℕ	
		Journal	Journal ID	ℕ	
		NbAut	Number of authors	ℕ	
		Apos	Position of the focal author in the list of article authors	ℕ	
		ACitBF	Yearly number of article citations at author's HHMI appointment	ℕ	yes
		ACitAF	Yearly number of article citations after author's HHMI appointment	ℕ	yes
		IF	Impact factor of the journal	ℝ	
		HIF	Journal has high impact factor	0/1	
		LIF	Journal has low impact factor	0/1	
		Novelty	Novelty of the article (i.e., mean age of the keywords)	ℝ	
		Novel	Article is novel	0/1	
		NotNovel	Article is not novel	0/1	
		Recombination	Level of recombination of the article (i.e., proportion of out-of-field literature cited)	ℝ	
		Recombinant	Article is recombinant	0/1	
		NotRecombinant	Article is not recombinant	0/1	
	Unobserved	ArtQ	Quality of the article	ℝ	
		UArtQ	Uncertainty about the quality of the article	ℝ	

Table 4. List of regressions

Regression	**Description**
r_1	Based on the full sample. Independent variables include all the observed variables listed in Table 3
r_2	Like r_1, but the sample includes only articles published up to 1 year before the HHMI appointment
r_3	Like r_1, but the sample includes only articles published 2 years before the HHMI appointment
r_4	Like r_1, but the sample includes only articles published 3 to 10 years before the HHMI appointment
r_5	Like r_2, but estimating a different effect of HHMI when HIF is true or false
r_6	Like r_2, but estimating a different effect of HHMI when Novel is true or false
r_7	Like r_2, but estimating a different effect of HHMI when Recombinant is true or false
r_8	Like r_2, but estimating a different effect of HHMI when Hwellcited is true or false
r_9	Like r_2, but estimating a different effect of HHMI when Hyoung is true or false
r_a	Hypothetical regression where the direct Matthew effect on Status is estimated
r_b	Hypothetical regression where the effects of HIF, Novel, and Recombinant on UArtQ are estimated
r_c	Hypothetical regression where the effect of UArtQ on ACitAF is estimated
r_d	Hypothetical regression where the effects of Hwellcited and Hyoung on Status are estimated
r_e	Hypothetical regression where the effect of Status on ACitAF is estimated
r_f	Hypothetical regression where the effect of Status on HHMI is estimated