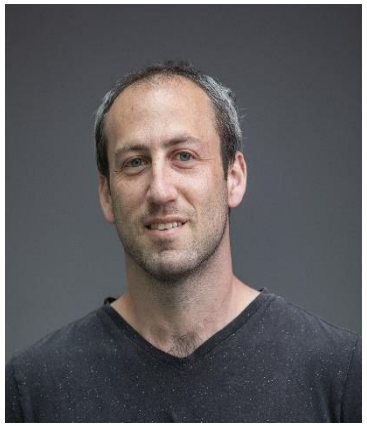
Is this correct? Let's Check!

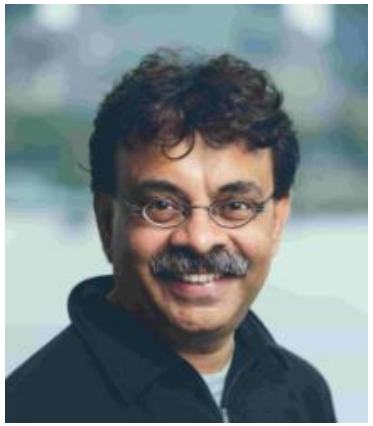
Omri Ben-Eliezer

Dan Mikulincer



Elchanan Mossel

Madhu Sudan



Is societal knowledge robust?

- Why ask this question?
  - Builds on error-prone processes
    - Collecting Data
    - Analyzing it
    - Combining results
- Last is especially problematic/interesting: Knowledge is cumulative!!
  - Accumulation can be very bad for errors!!!!!
- There must exist error-correcting processes
  - What are they? How do they work? How well do they work?

Lebesgue's Mistake

- In 1904 Lebesgue proved the following theorem:

  "A projection of a measurable set is measurable"

Sur les fonctions représentables analytiquement ;

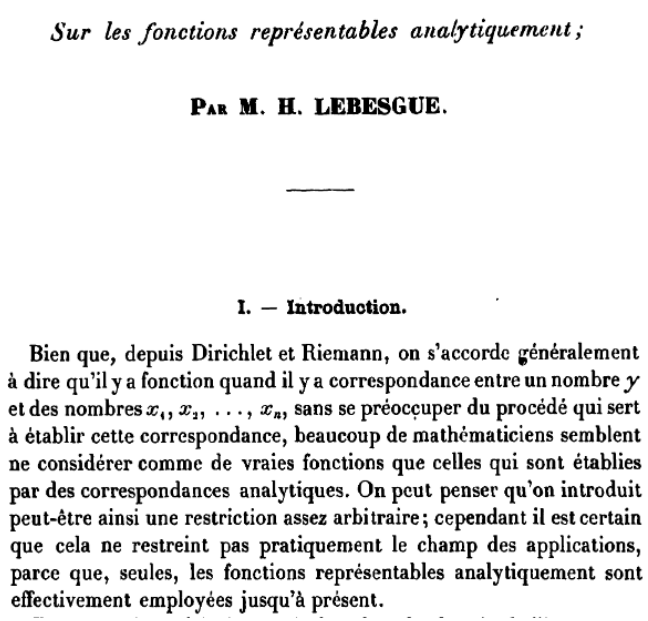
PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire ; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.

Lebesgue's Mistake

- According to Google Scholar the paper has 303 citations.

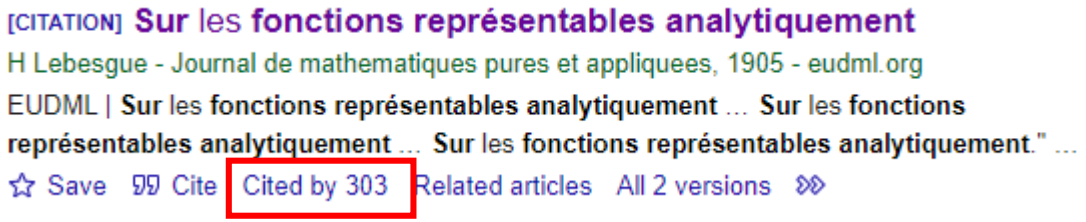


Sur les fonctions représentables analytiquement ;

PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire ; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.



[CITATION] **Sur les fonctions représentables analytiquement**

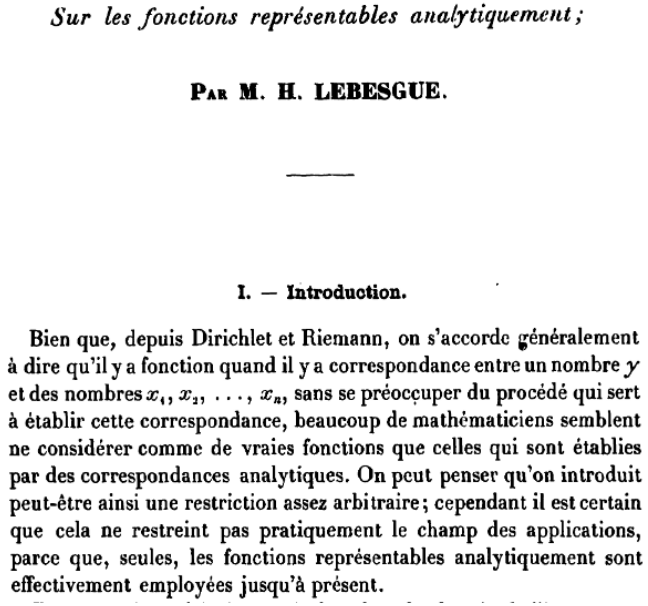
H Lebesgue - Journal de mathematiques pures et appliquees, 1905 - eudml.org

EUDML | **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement**." ...

☆ Save Cite Cited by 303 Related articles All 2 versions »

Lebesgue's Mistake

- According to Google Scholar the paper has 303 citations.
- Some citations prior to 1917.

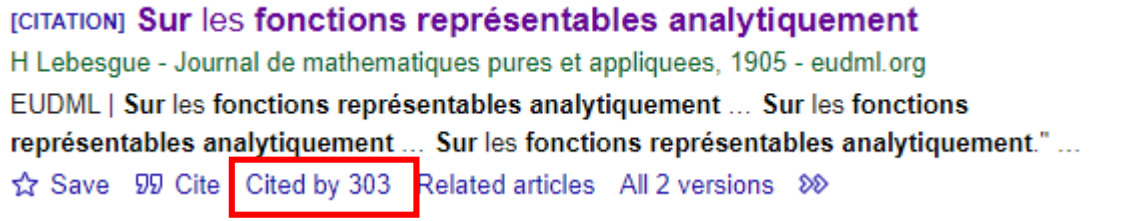


Sur les fonctions représentables analytiquement;

PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.



[CITATION] **Sur les fonctions représentables analytiquement**

H Lebesgue - Journal de mathematiques pures et appliquees, 1905 - eudml.org

EUDML | **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement**." ...

☆ Save 99 Cite Cited by 303 Related articles All 2 versions »

Sulla rappresentazione analitica delle funzioni di piÙ variabili reali

Leonida Tonelli

Rendiconti del Circolo Matematico di Palermo **29**, 1–36 (1910) Cite this article

References

2) *Sur Us fonctions représentables analytiquement* [Journal de Mathématiques pures et appliquées, VIe série, I (1905), pp. 139–216].

Lebesgue's Mistake

- According to Google Scholar the paper has 303 citations.
- Some citations are from before 1917.
- In 1917 Suslin discovered a counterexample:

  "There exists a projection of measurable set which is not mesurable"
- Happy ending: The field of "Descriptive set theory" was born.
- **Did the mistake propagate?**

Sulla rappresentazione analitica delle funzioni di piÙ variabili reali

Leonida Tonelli

Rendiconti del Circolo Matematico di Palermo **29**, 1–36 (1910) Cite this article

References

2) *Sur Us fonctions représentables analytiquement* [Journal de Mathématiques pures et appliquées, VIe série, I (1905), pp. 139–216].

Today

Question

Can we guarantee that the effects caused by a single error do not propagate?

Sur les fonctions représentables analytiquement ;

PAR M. H. LEBESGUE.

I. — Introduction.

Bien que, depuis Dirichlet et Riemann, on s'accorde généralement à dire qu'il y a fonction quand il y a correspondance entre un nombre y et des nombres $x_1, x_2, \ldots, x_n$, sans se préoccuper du procédé qui sert à établir cette correspondance, beaucoup de mathématiciens semblent ne considérer comme de vraies fonctions que celles qui sont établies par des correspondances analytiques. On peut penser qu'on introduit peut-être ainsi une restriction assez arbitraire ; cependant il est certain que cela ne restreint pas pratiquement le champ des applications, parce que, seules, les fonctions représentables analytiquement sont effectivement employées jusqu'à présent.

[CITATION] **Sur les fonctions représentables analytiquement**

H Lebesgue - Journal de mathematiques pures et appliquees, 1905 - eudml.org

EUDML | **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement** ... **Sur les fonctions représentables analytiquement**." ...

☆ Save Cite Cited by 303 Related articles All 2 versions »

Sulla rappresentazione analitica delle funzioni di piÙ variabili reali

Leonida Tonelli

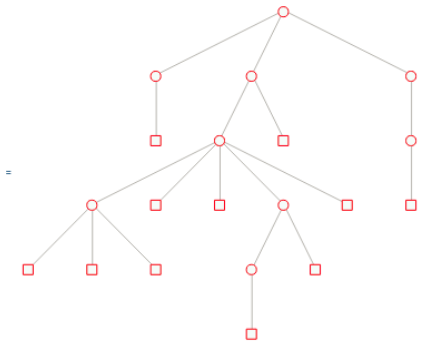
Rendiconti del Circolo Matematico di Palermo **29**, 1–36 (1910) Cite this article

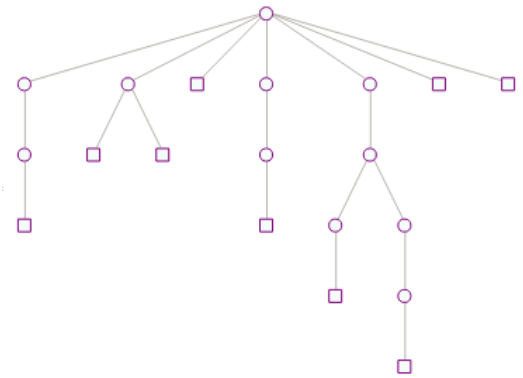
References

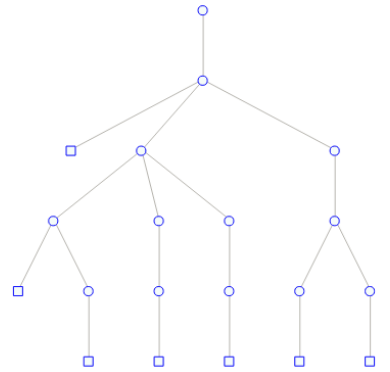
2) *Sur Us fonctions représentables analytiquement* [Journal de Mathématiques pures et appliquées, VIe série, I (1905), pp. 139–216].

Cumulative Knowledge Process

- In the paper we model the process of accumulating knowledge.
- Main properties:
  - New “units of knowledge” build upon previous units.
  - Errors are sometimes introduced and may propagate forward.
  - Errors can be checked and removed from the process.
- We study **structural properties** of the process.



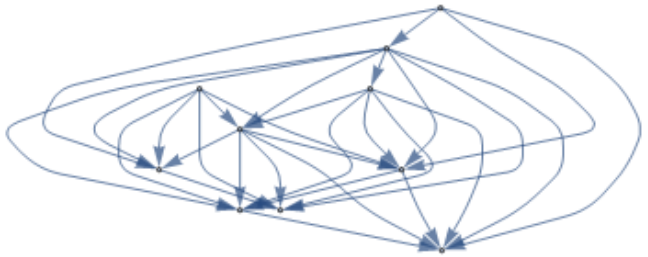




The Model

Representation of Knowledge

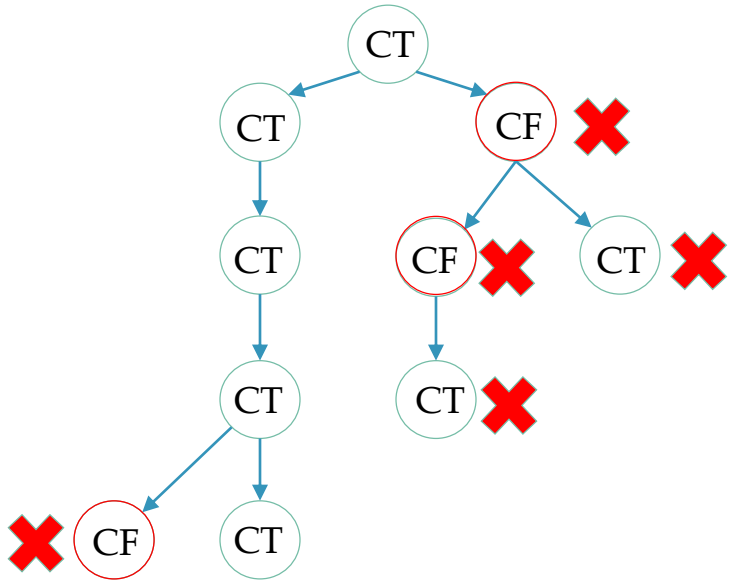
- **Ideally**: Knowledge is stored as Directed Acyclic Graph (DAG).
  - Vertices represent units of knowledge
  - Edges represent dependence or “inherited knowledge”.
  - E.g. a paper cites several papers.



- Would require a proper model of knowledge clustering.
- **Simplified notion:** Knowledge is represented as a tree.

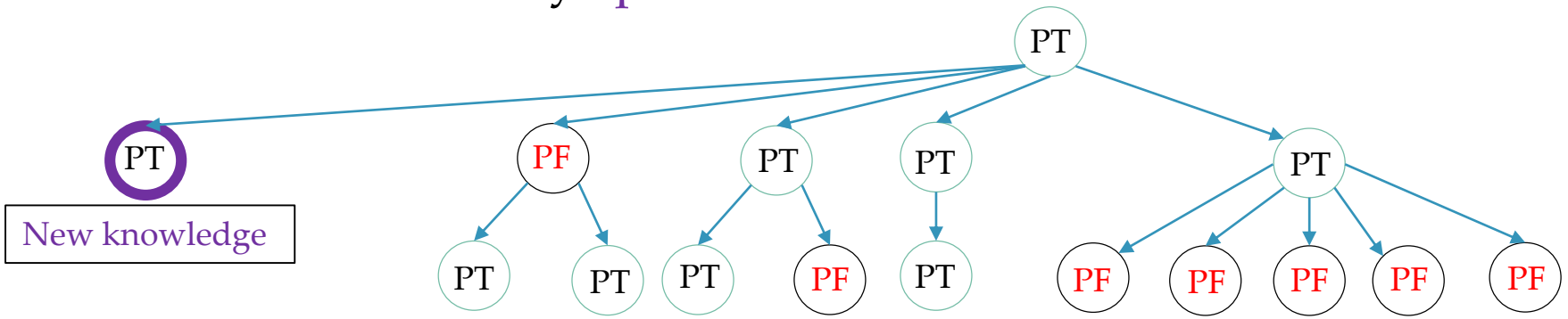
The Model

- **The Knowledge ~~DAG~~ Tree**: In the Cumulative Knowledge Process (CKP) knowledge units are modeled as a tree.
- A **node** represents a single "unit of knowledge" and **edges** represent the relation of "building upon existing knowledge"
- Each node has:
  - A hidden state conditionally true(CT)/conditionally false(CF)
  - A public state proclaimed true(PT)/proclaimed false(PF)
  - A node is considered to hold true knowledge
    if **all ancestors are CT**



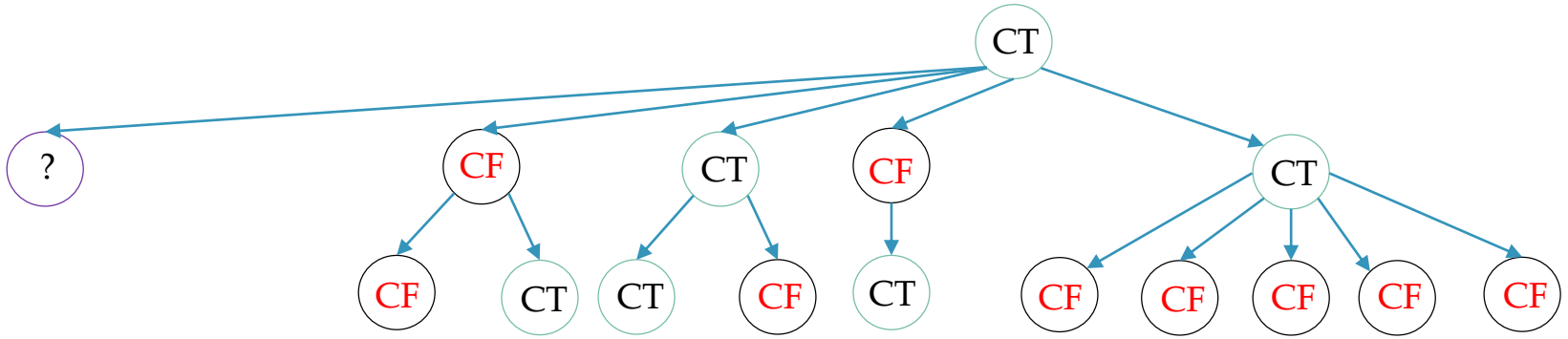
Accumulating Knowledge

- At each time $t \geq 0$, we have a knowledge tree T_t with associated labels.
- At time $t + 1$ a new node is added to the tree, by choosing a **random** proclaimed true (PT) parent.
- Parents are chosen according to the preferential attachment model.
  - **The more PT children a node has the more likely it is to generate new knowledge.**
- A new node is always proclaimed true



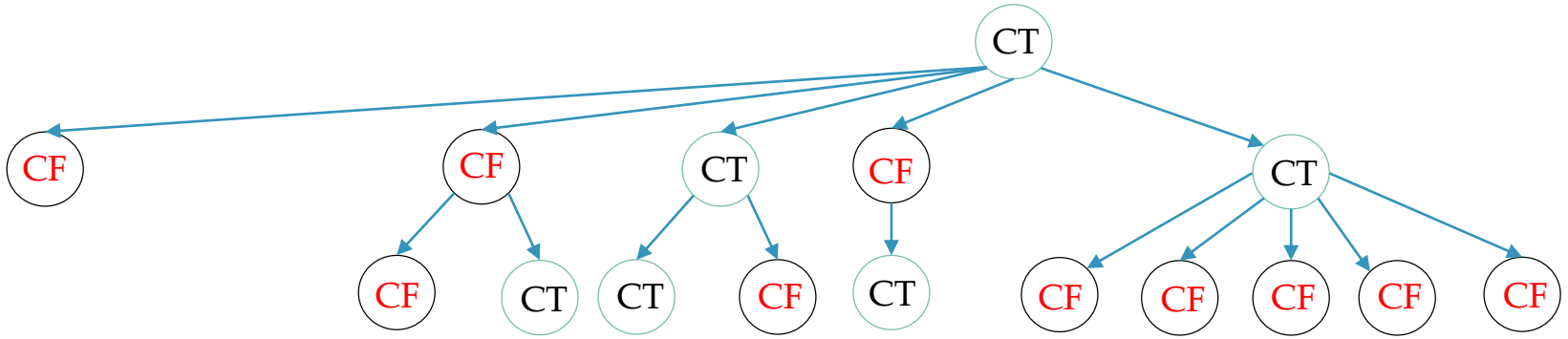
Injection of Errors

- Recall: nodes also have hidden states.
- The hidden label of a new node is determined randomly:
  - Parameter ε: with probability ε the new node is CF and otherwise CT.



Injection of Errors

- Recall: nodes also have hidden states.
- The hidden label of a new node is determined randomly:
  - Parameter ε: with probability ε the new node is CF and otherwise CT.

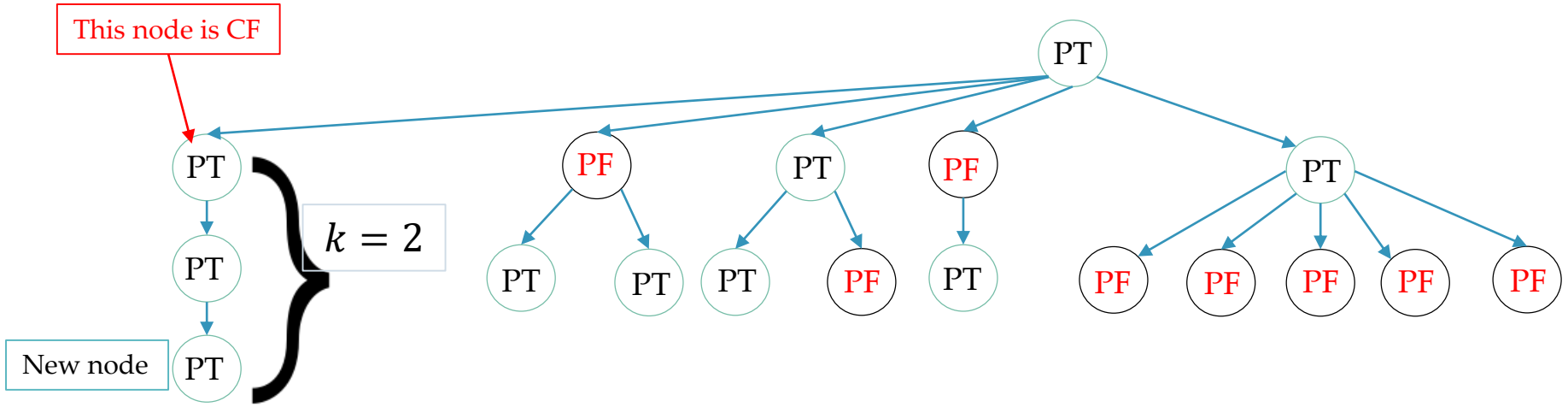


Aside: Probabilistic Models

- Quote from unknown source*
  - "All models are wrong. Some are useful"

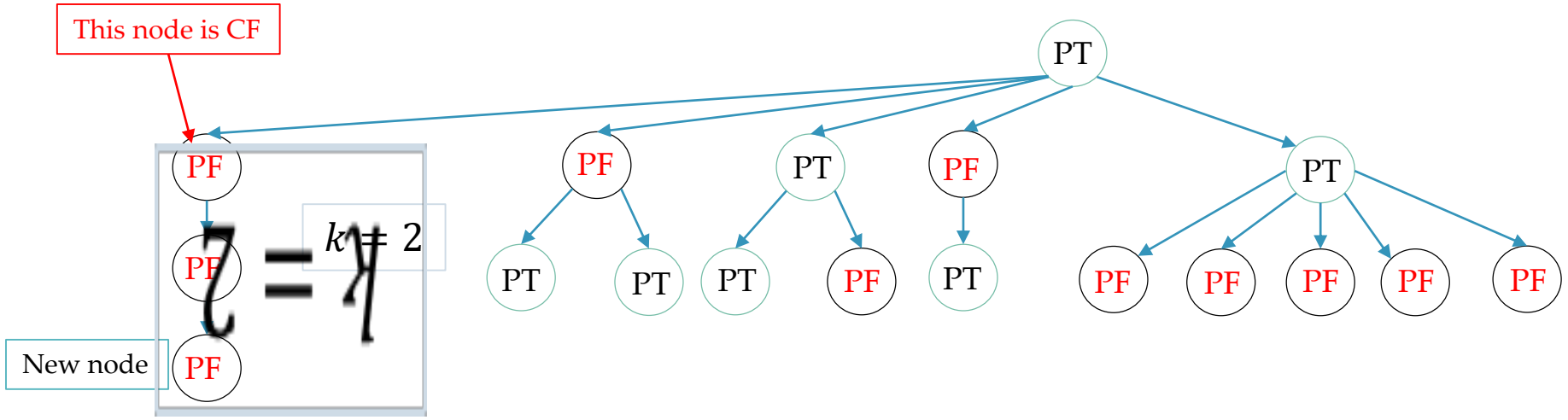
Checking for Errors

- Checks may be performed whenever a new node is added.
  - Parameter p: a node preformed a check with probability p.
- Checks are performed by ascending the tree.
  - Parameter k: the number of levels to be checked.



Checking for Errors

- Checks may be performed whenever a new node is added.
  - Parameter p: a node preformed a check with probability p.
- Checks are performed by ascending the tree.
  - Parameter k: the number of levels to be checked.
- If a CF or PF node is encountered, the public state of the **entire path** changes to proclaimed false.



The Model - Summary

- **Tree**: knowledge is represented by a tree.
- **Preferential attachment**: nodes of high degree are more influential.
- **Errors:** errors are sometimes introduced when new knowledge is created
- **Checks:** checks are sometimes preformed when new knowledge is introduced.
- **Error correction:** when a node is verified to be faulty, the error is announced and the node is effectively eliminated.

With the parameters $\varepsilon, p,$ and k the model is called the $(\boldsymbol{\varepsilon, p, k}) - \boldsymbol{CKP}$.

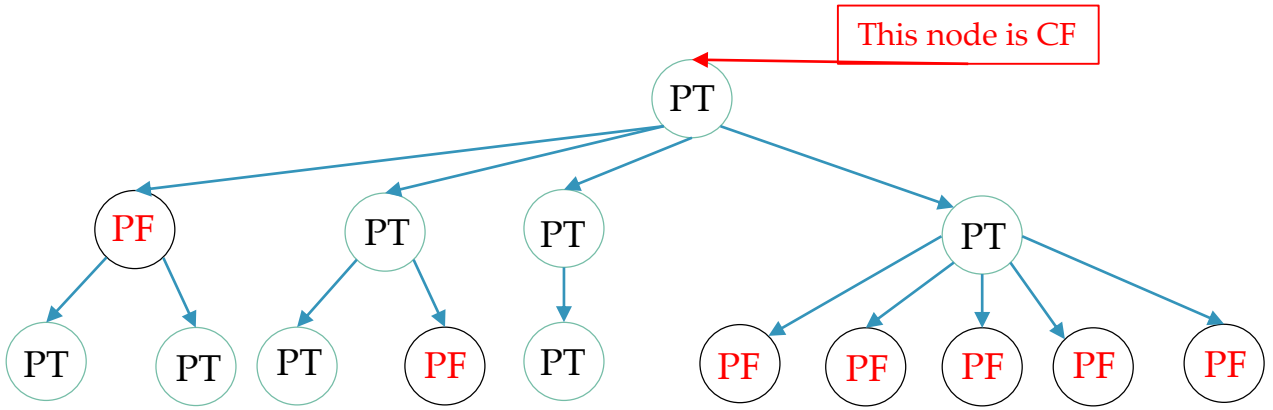
Results

Error Effects

Question

Can we guarantee that the effects caused by a single error do not propagate?

- **Effects caused by a single error:** subtree rooted at a CF node.

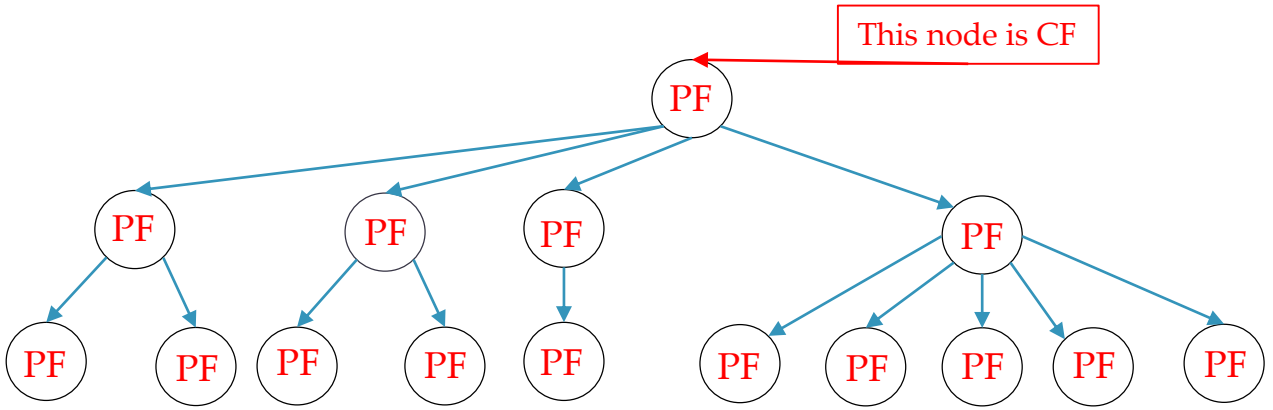


Error Effects

Question

Can we guarantee that the effects caused by a single error do not propagate?

- **Effects caused by a single error:** subtree rooted at a CF node.
- **Elimination of errors (observation):** if, at some time, all nodes in a subtree are marked PF, the subtree is effectively eliminated.



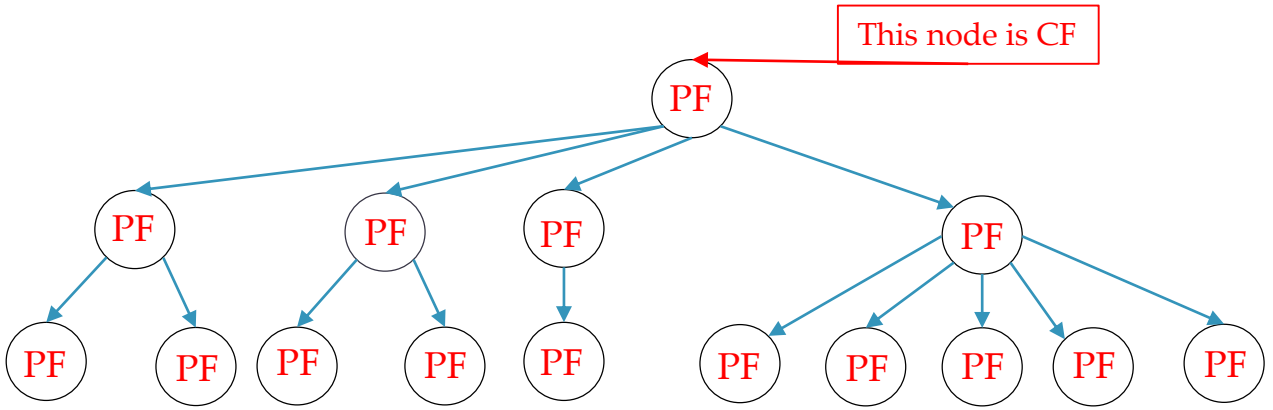
Error Effects

Question

Can we guarantee that the effects caused by a single error do not propagate?

Definition

- If every subtree rooted at a CF node is eliminated with probability 1, we say that the **error effects are completely eliminated.**
- Otherwise, the **error effects survive with positive probability.**

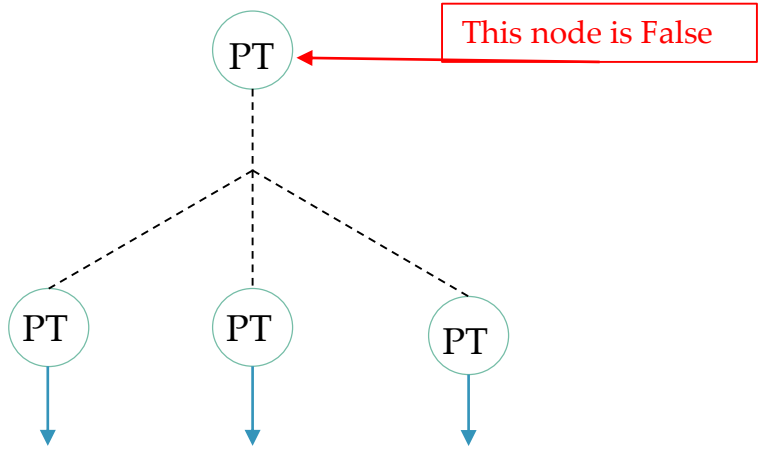


First Result – Depth Matters

Theorem 1

If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survives with positive probability.

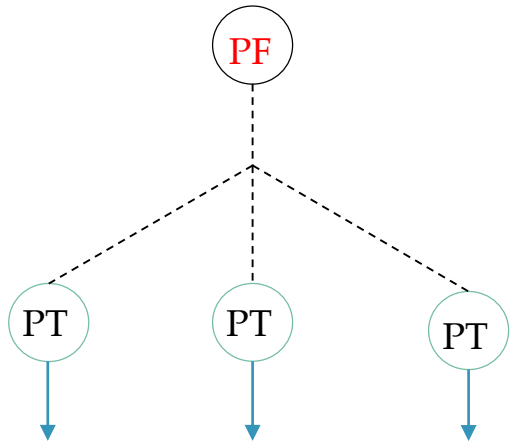
- Main idea: couple the $\boldsymbol{CKP}$ with a branching process.



First Result – Depth Matters

Theorem 1
If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survive with positive probability.

- Main idea: couple the ***CKP*** with a branching process.
- We show that when $k = 2$, by the time an erroneous node is proclaimed false it will effectively create many new components.



First Result – Depth Matters

Theorem 1

If $\boldsymbol{k = 2}$, then for any $\boldsymbol{p < 1}$, the error effects in the $\boldsymbol{(\varepsilon, p, 2) - CKP}$ survive with positive probability.

Conclusion:

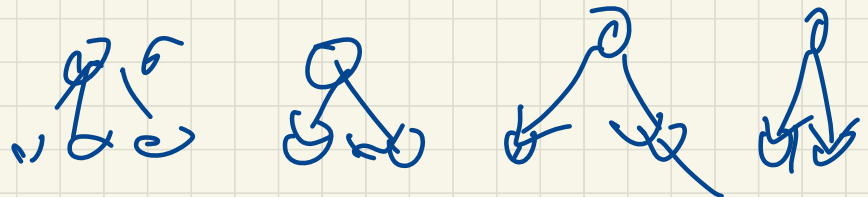
To guarantee that error effects are completely eliminated shallow checks are not enough!

Another Result: if $p < \frac{1}{4}$, then $k = \infty$

is not enough to eliminate error effects.

($t = 0$; root is CF)

$p_k > \dots$ large $\Rightarrow$ means



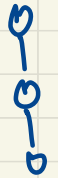
Analysis

Some potential function

$\Rightarrow$ grows in expectation $\Rightarrow$ process lives for ever

(# leaves
+ # unrooted components)

$\Rightarrow$ shrinks in expectation $\Rightarrow$ process dies



exponential potential in depth

Second Result– Checking Matters

Theorem 2

For any $\boldsymbol{k \geq 4}$, and $\boldsymbol{\varepsilon < 1}$, there exists $p_0 \in (0,1)$ such that:

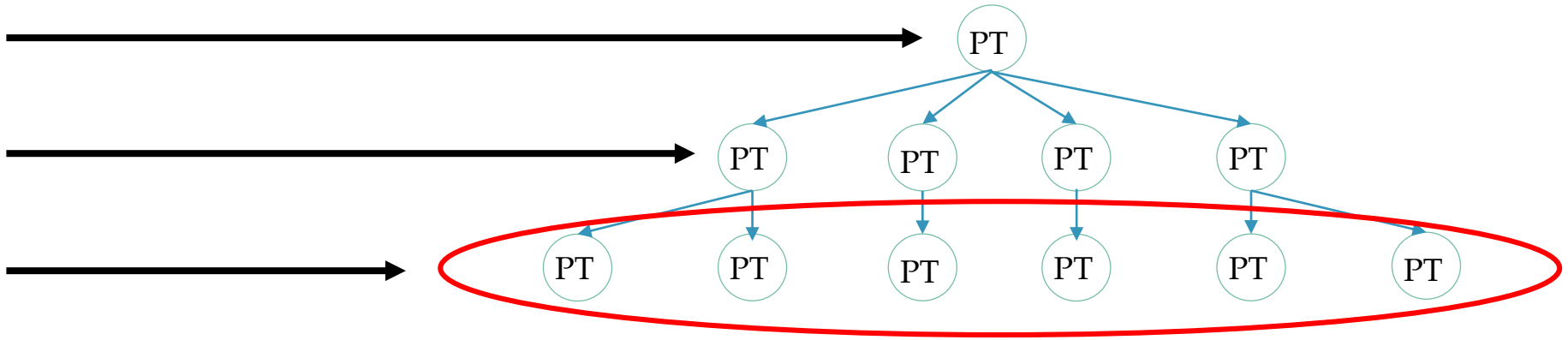
- If $p > p_0$ the error effects in the $(\boldsymbol{\varepsilon, p, k}) - \boldsymbol{CKP}$ are completely eliminated.
- If $p < p_0$ the error effects in the $(\boldsymbol{\varepsilon, p, k}) - \boldsymbol{CKP}$ survive with positive probability.

Conclusion:

When the checking procedure is not too shallow, there is a minimal amount of effort to invest in checking to guarantee the elimination of error effects.

Second Result– Main Ideas

- Proof of Theorem 2 is based on a (sub\super-)martingale analysis.
- We consider some observables in the process and identify regimes in which they increase or decrease in expectation.
- Examples:
  - Number of proclaimed true leaves in the tree.
  - Distribution of depths in proclaimed true subtree.



Further Results

- With a refined analysis we also consider other structural properties.
- When the *error effect survives*:
  - Identify parameters which ensure that **false components are sublinear**.
  - Also **control the size** of the components.
- When the *error effect is completely eliminated*:
  - Identify parameters which also ensure that **proportion of false nodes** in the tree is always **at the noise level**.

Future Directions

- **The mysterious case of $k = 3$**:
  - Can the error effect be eliminated when only preforming depth 3 checks?
- **Phase transitions**:
  - Determine the value of the critical probability p_0.
- **More general models**:
  - Can similar results be obtained for DAGS, instead of tree?
  - Will require to define an appropriate preferential attachment model on DAGs, which allows "similar knowledge" units to cluster.

Thank you!