



Lazy Prices Replication:
Do investors ignore Repetitive Disclosures

What did the paper say & What we are trying to do?

Paper:

Some firms copy-and-paste their reports every year. Investors don't seem to notice — and those "lazy" firms often outperform. The paper says:

repetition = return opportunity.

The paper argues that the investors do not fully process this, so they usually underreact. This means that the "repetitive" firm may be undervalued in the short term. As a result, these firms often earn higher future return.

Our Project:

We tested this idea using real 10-Ks and 10-Qs and stock data. → 5 reports per year

We built a signal that measures **how much firms repeat themselves**, and used it to build **long-short trading portfolios**.

Data Overview

SEC Filings:

Thousands of annual and quarterly reports (10-Ks/10-Qs) downloaded from the SEC EDGAR database. We use the dictionary that ML build.

Stock Returns:

Monthly returns for all U.S. public companies from CRSP, adjusted for dividends and splits.

Open Asset Pricing Benchmark:

BM and Mom12m signals from OAP used to benchmark our portfolio's performance.

Custom ID Mapping:

We built a unique mapping to link company identifiers (PERMNO, CIK, ticker) across all datasets—this made accurate merging possible.

Step 1: Parsing 10-K Filing Dictionaries

- We started with a dataset of SEC 10-K filings converted into word frequency vectors
 - each row representing one firm's filing.
- To manage the large file size
 - **chunks of 10,000 filings** at a time.
- From each chunk
 - CIK
 - Filing_data
 - Word count features
- Data Cleaning
 - Filing missing word frequencies with 0
 - Converting filing date to datetime format

chunk_id	CIK	filing_date	word_1	word_2	word_3
0	100001	2011-03-15	3	1	0
0	100002	2012-07-29	6	5	1
0	100003	2010-01-10	2	3	2
1	100101	2013-05-03	5	0	3
1	100102	2011-11-20	1	2	7
2	100201	2015-08-17	4	3	2

Step 2: Calculating similarity score






- We checked each firm's 10-k one year at a time
 - Think? How similar is this year's filing to last year's?
- Decide to use cosine similarity
 - A number between 0 and 1
- How to interpret?
 - Close to 1 = the firm repeated itself
 - Lower score = the firm changed its wording

Firm	Year	Filing (Word Vector)	Cosine Similarity to Previous Year
A	2019	[3, 2, 0, 4, 1]	— (first filing)
A	2020	[3, 2, 0, 4, 1]	1.00 (identical to 2019)
B	2019	[1, 5, 2, 0, 4]	—
B	2020	[7, 1, 2, 0, 3]	0.52 (some wording changes)

- The scores tells us how much a firm is "repeating" itself
- We saved one similarity score per firm, per year
 - This become the signal we used to build trading strategies.

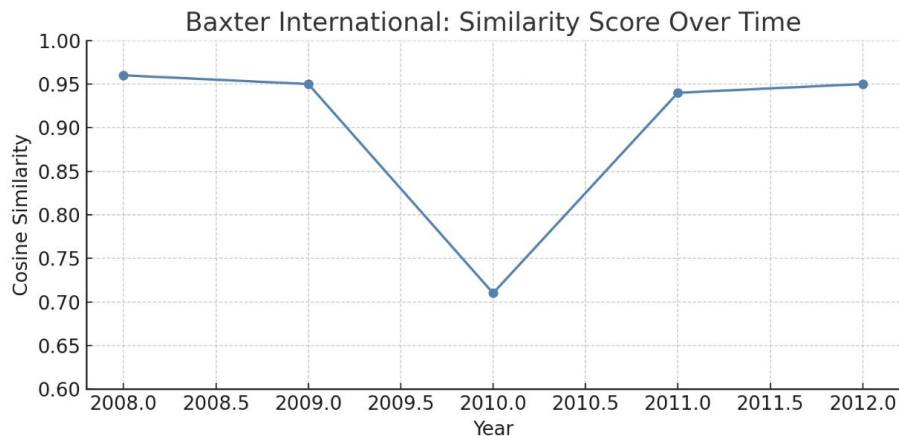
Step 3: Mapping Firms and Merging with Stock Returns

- We had textual similarity data
 - Indexed by SEC CIKs
- We had stock return data
- These need to be matched correctly to build trading signals
- We create a custom CIK-PERMNO mapping
 - Any mismatch would break the analysis
- After merge, we aligned monthly returns with similarity score

Ticker	CIK	PERMNO	Match Status	Notes
AAPL	0000320193	14593	 Matched	Clean match via ticker and CIK
MSFT	0000789019	10107	 Matched	Ticker and CIK aligned successfully
XYZB	0001234567	—	 Missing	No match in CRSP (delisted stock)
GOOG	0001652044	13141	 Matched	Needed lowercase + padding fix
ABCA	0009999999	—	 Dropped	Invalid/placeholder ticker

Step 4: Exploring the Similarity Signal

- Some firms had scores close to **1** every year → they reused the same language
- Others had much **lower scores**, suggesting big changes in their disclosures
- We used examples like **Baxter International** to see real drops in similarity
- Keys:
 - Capture real disclosure behavior
 - It varies across firms and time → useful for portfolio construction



Step 5: Building Portfolios Overall Logic

- For each month, firms are stored by their similarity score
- Divided into quintiles
 - Top 20% → highest score → Long
 - Bottom 20% → Lowest score → Short
- Long-Short = Mean (Long Leg Returns) - Mean(Short Leg Returns)

Firm	Month	Similarity Score	Quintile	Action	Return (%)
A	Jan	0.95	Q5	Long (Buy)	3.0
B	Jan	0.88	Q5	Long (Buy)	2.5
C	Jan	0.72	Q3	Hold	1.0
D	Jan	0.41	Q2	Hold	0.5
E	Jan	0.20	Q1	Short (Sell)	-1.5
F	Jan	0.15	Q1	Short (Sell)	-2.0

Step 5: Building Portfolios

Winsorized Portfolio

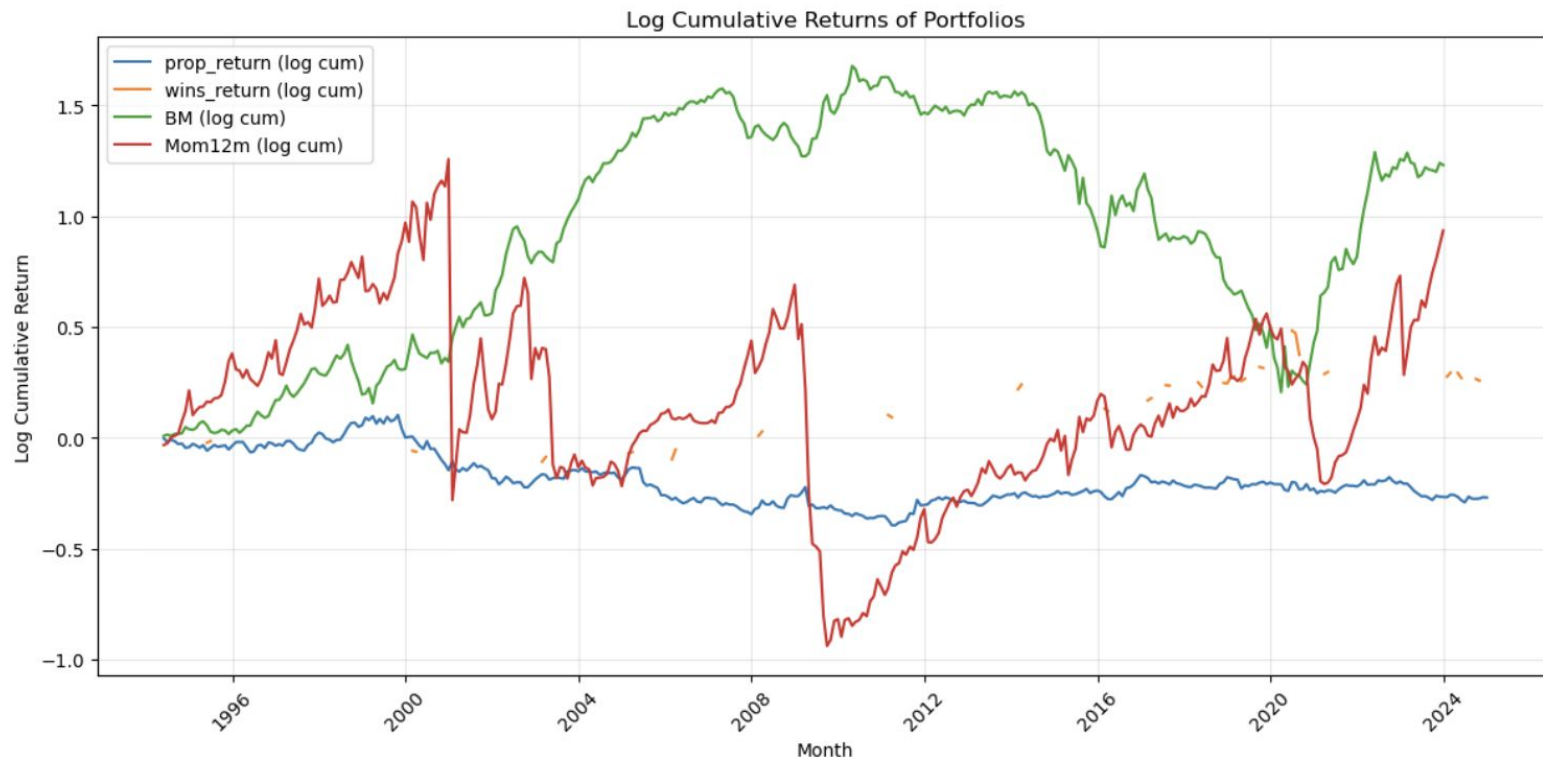
- Only included months
 - We successfully computed similarity score
- Used a left merge for returns
 - Making sure we only keep the months with similarity scores
- Reason Why we did like this
 - This portfolio gave us more stable and positive average return

Propagated Portfolio

- Make the similarity signal apply to every month
 - Fills forward the latest similarity score into future months
- Used a right merge to keep all available return data

Firm	Month	Similarity Score	Signal Used?	Quintile	Action	Return (%)
A	Jan	0.90	✓ Yes	Q5	Long (Buy)	2.0
B	Jan	NaN	✗ Propagate from Dec	Q2*	Wrong bin	0.5
C	Jan	NaN	✗ Propagate from Dec	Q5*	Wrong bin	-1.0
D	Jan	0.20	✓ Yes	Q1	Short (Sell)	-1.5
E	Jan	0.18	✓ Yes	Q1	Short (Sell)	-2.2

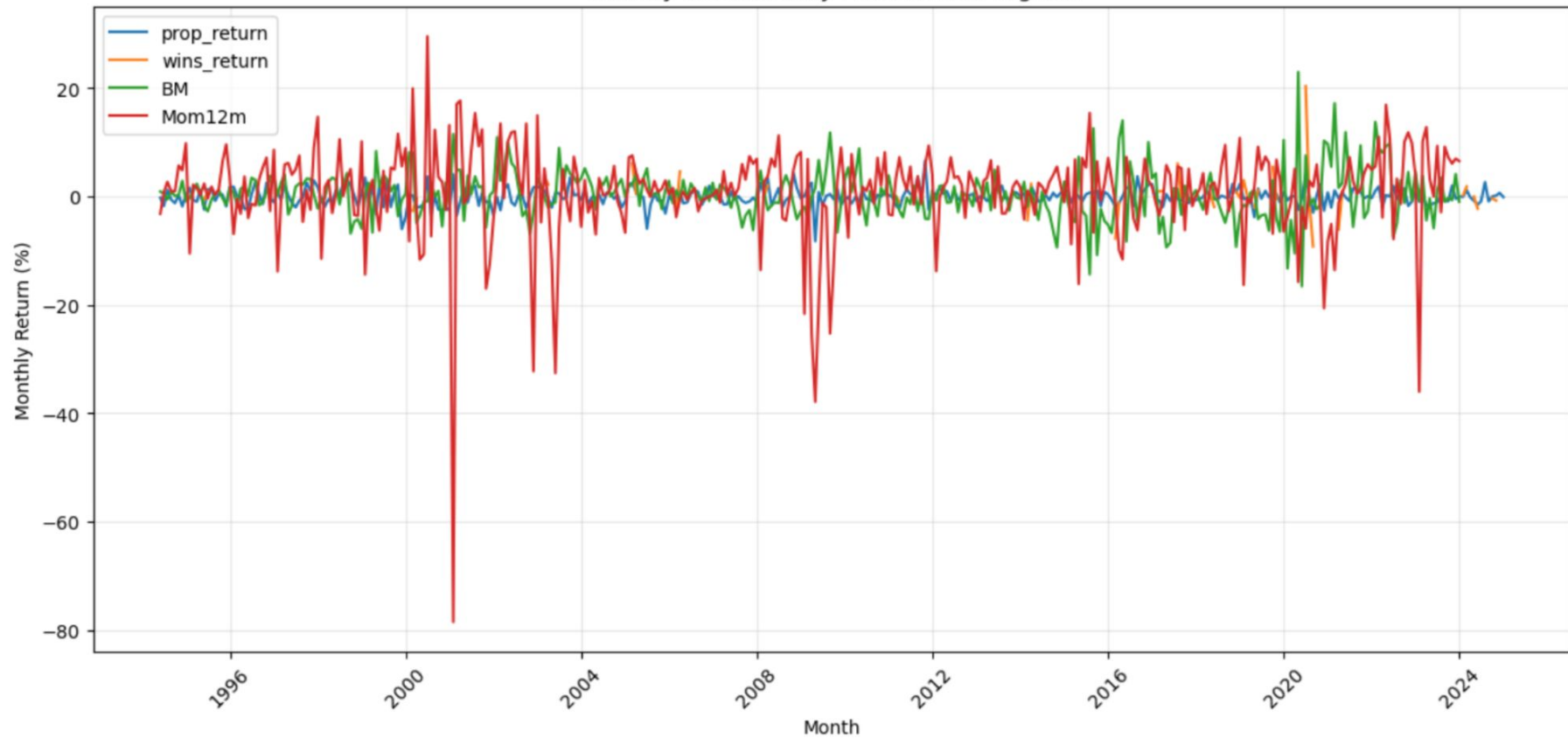
Step 6: Result



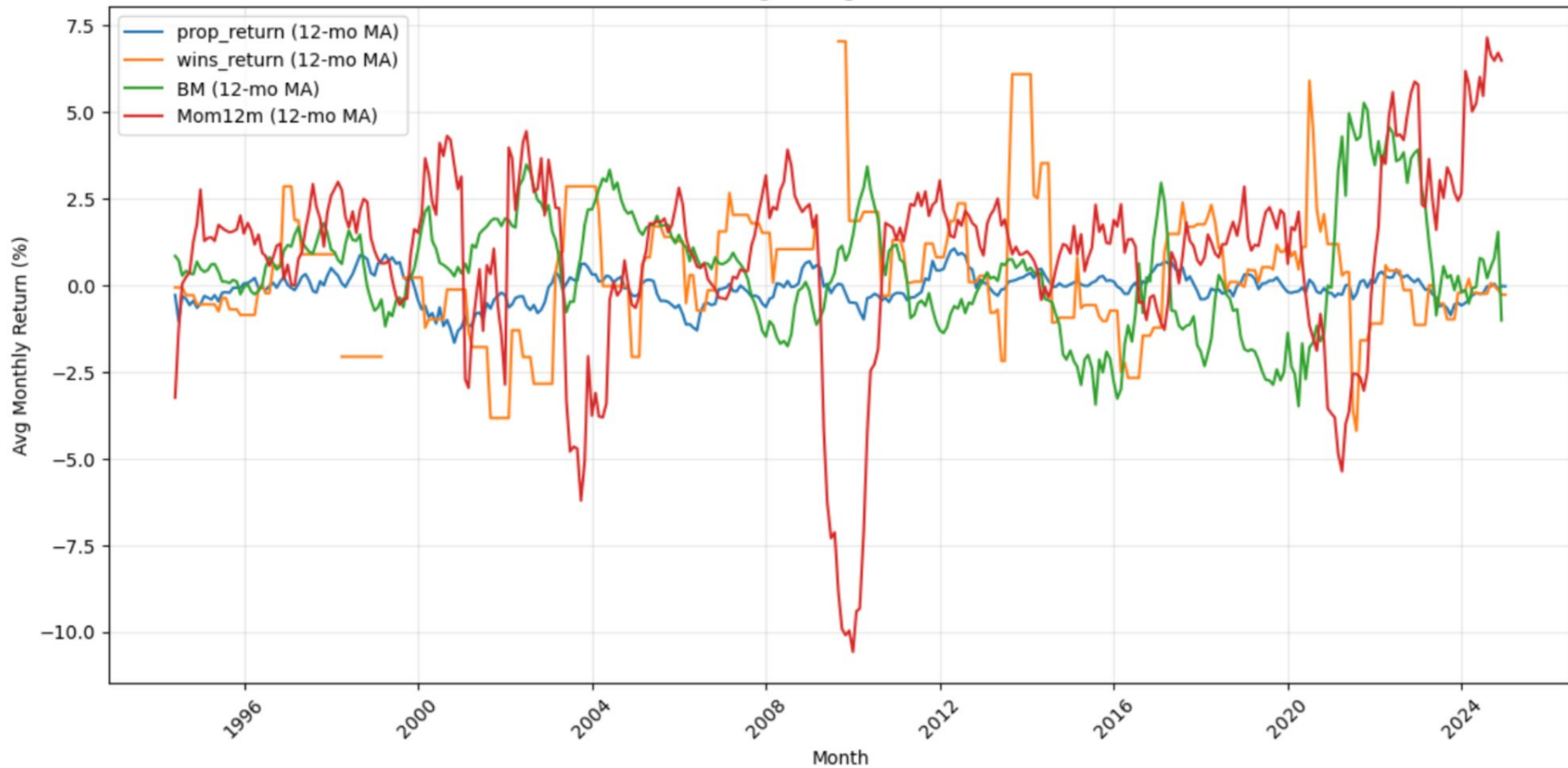
Cumulative Returns: Lazy Prices vs OAP Signals



Monthly Returns: Lazy Prices vs OAP Signals



12-Month Moving Averages of Portfolio Returns



Step 6: Result

- **Orange line (winsorized strategy)** shows steady growth → signal works when implemented carefully
- **Blue line (propagated strategy)** is flat/declining → performance suffers from stale or missing similarity scores
- Confirms: **signal quality and freshness are essential** for return predictability
- Supports the **Lazy Prices hypothesis**: repetitive disclosures may be underreacted to by the market
- Highlights the importance of **data engineering** in empirical finance research

Thanks

