On the Origins of Memes by Means of Fringe Web Communities

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil.

In Proceedings of International Measurement Conference (IMC) ’18. ACM, 15 pages.

情報工学科 寺岡研究室

安森 涼 61619027

背景

- Web の台頭
  - 考え, 文化, 画像, 映像などが今までにない速度で拡散
- 政治思想や信条の拡散に**ミーム (meme)** を利用
  - ミーム : インターネットで流行する画像や言い回し
- 攻撃的なミーム特定の需要が増加
  - e.g., 人種差別的, 政治的

SNS ごとにミームの分布, 拡散の傾向の解釈が必要

通常

人種差別的

政治的

関連研究

- Twitter のハッシュタグからミームの人気度調査 [1]
- Facebook 間の伝搬の過程で新しいミームが生まれることを証明 [2]
- 4chan, Reddit で口汚い言葉を用いた投稿の検知 [3]

- **検証対象外の SNS データセットへの適用は不可能**
  - **複数 SNS のミームの意味付けが必要**
    → ハッシュ化した画像のクラスタリングを実施
- **ミームの発生元となる SNS の未調査**
  - **複数 SNS 間のミームの伝搬の検知が必要**
    → 伝搬性をもつ事象を表現できるモデルを作成

[1] L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting Successful Memes Using Network and Community Structure. In ICWSM, 2014.
[2] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng. Information Evolution in Social Networks. In WSDM, 2016.
[3] E. Chandrasekharan, et al., The bag of communities: Identifying abusive behavior online with preexisting Internet data. In CHI, 2017.

データセット

- 調査対象の SNS

※ 2016/07/01 から 2017/07/31 のデータを利用

- **Mainstream : 悪意のないミームが拡散されやすい SNS**

ウェブサイト名	概要
Twitter	140 字以内の短い記事を投稿し合うサイト
Reddit	ニュース記事, 画像やテキストの投稿サイト

- **Fringe : 悪意を含むミームが拡散されやすい SNS**

ウェブサイト名	概要
The_Donald (T_D)	Reddit のトランプについてのチャンネル
/pol/	匿名の掲示板 4chan の政治チャンネル
Gab	ほぼ規制がない, 言論の自由を尊重するサイト

- ミームのまとめサイト (キュレーションサービス)

ウェブサイト名	概要
Know Your Meme (KYM)	ミームに対するメタデータを供給

提案手法 1｜複数 SNS のミームを意味付け

1. ミームの一部をハッシュ化
2. クラスタリング
3. クラスタごとに Medoid を決定
4. クラスタとメタデータを結び付け
5. 他のデータセットへの適用

1. 64 bit ハッシュ値を利用
   - 似た画像は近い値を保持
2. ハミング距離準拠の
   アルゴリズムを利用
3. Medoid : クラスタ内の
   自身以外の画像との距離の
   総和が最小となる画像
4. ハッシュ化した
   KYM データを利用
5. ハッシュ化した
   全てのデータを利用

入力が画像のため任意の SNS に対応可能

評価 1-1｜SNS 毎に投稿されるミームの割合

- 人種差別的なミーム (図中赤) は Fringe で上位
  - ‘racist’, ‘antisemitism’ などのタグを持つもの
- 悪意のないミームは Mainstream で上位
- 政治的なミーム (図中黄) はどこにでも存在
  - ‘politics’, ‘trump’, ‘clinton’ などのタグを持つもの

Fringe　　Mainstream

/pol/		Gab		Twitter		Reddit	
Entry	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts(%)**	**Entry**	**Posts (%)**
Feels Bad Man/Sad Frog	64,367 (4.9%)	Jesusland **(P)**	454 (1.6%)	Roll Safe	55,010 (5.9%)	Manning Face	12,540 (2.2%)
Smug Frog	63,290 (4.8%)	Demotivational Posters	414 (1.5%)	Evil Kermit	50,642 (5.4%)	That’s the Joke	7,626 (1.3%)
Happy Merchant **(R)**	49,608 (3.8%)	Smug Frog	392 (1.4%)	Arthur’s Fist	37,591 (4.0%)	Feels Bad Man/ Sad Frog	7,240 (1.3%)
Apu Apustaja	29,756 (2.2%)	Based Stickman **(P)**	391 (1.4%)	Nut Button	13,598 (1,5%)	Confession Bear	7,147 (1.3%)
Pepe the Frog	25,197 (1.9%)	Pepe the Frog	378 (1.3%)	Spongebob Mock	11,136 (1,2%)	This is Fine	5,032 (0.9%)
Make America Great Again **(P)**	21,229 (1.6%)	Happy Merchant **(R)**	297 (1.1%)	Reaction Images	9,387 (1.0%)	Smug Frog	4,642 (0.8%)
Angry Pepe	20,485 (1.5%)	Murica	274 (1.0%)	Conceited Reaction	9,106 (1.0%)	Roll Safe	4,523 (0.8%)
Bait this is Bait	16,686 (1.2%)	And Its Gone	235 (0.9%)	Expanding Brain	8,701 (0.9%)	Rage Guy	4,491 (0.8%)
I Know that Feel Bro	14,490 (1.1%)	Make America Great Again **(P)**	207 (0.8%)	Demotivational Posters	7,781 (0.8%)	Make America Great Again **(P)**	4,440 (0.8%)
Cult of Kek	14,428 (1.1%)	Feels Bad Man/ Sad Frog	206 (0.8%)	Cash Me Ousside/Howbow Dah	5,972 (0.6%)	Fake CCG Cards	4,438 (0.8%)
Laughing Tom Cruise	14,312 (1.1%)	Trump’s First Order of Business **(P)**	192 (0.7%)	Salt Bae	5,375 (0.6%)	Confused Nick Young	4,024 (0.7%)
Awoo	13,767 (1.0%)	Kekistan	186 (0.6%)	Feels Bad Man/ Sad Frog	4,991 (0.5%)	Daily Struggle	4,015 (0.7%)
Tony Kornheiser’s Why	13,577 (1.0%)	Picardia **(P)**	183 (0.6%)	Math Lady/Confused Lady	4,722 (0.5%)	Expanding Brain	3,757 (0.7%)
Picardia **(P)**	13,540 (1.0%)	Things with Faces (Pareidolia)	156 (0.5%)	Computer Reaction Faces	4,720 (0.5%)	Demotivational Posters	3,419 (0.6%)
Big Grin / Never Ever	12,893 (1.0%)	Serbia Strong/Remove Kebab	149 (0.5%)	Clinton Trump Duet **(P)**	3,901 (0.4%)	Actual Advice Mallard	3,293 (0.6%)
Reaction Images	12,608 (0.9%)	Riot Hipster	148 (0.5%)	Kendrick Lamar Damn Album Cover	3,656 (0.4%)	Reaction Images	2,959 (0.5%)
Computer Reaction Faces	12,247 (0.9%)	Colorized History	144 (0.5%)	What in tarnation	3,363 (0.3%)	Handsome Face	2,675 (0.5%)
Wojak / Feels Guy	11,682 (0.9%)	Most Interesting Man in World	140 (0.5%)	Harambe the Gorilla	3,164 (0.3%)	Absolutely Disgusting	2,674 (0.5%)
Absolutely Disgusting	11,436 (0.8%)	Chuck Norris Facts	131 (0.4%)	I Know that Feel Bro	3,137 (0.3%)	Pepe the Frog	2,672 (0.5%)
Spurdo Sparde	9,581 (0.7%)	Roll Safe	131 (0.4%)	This is Fine	3,094 (0.3%)	Pretending to be Retarded	2,462 (0.4%)
Total	445,179 (33.4%)		4,808 (17.0%)		249,047 (26.4%)		94,069 (16.7%)

評価 1-2｜ミームの投稿数の推移

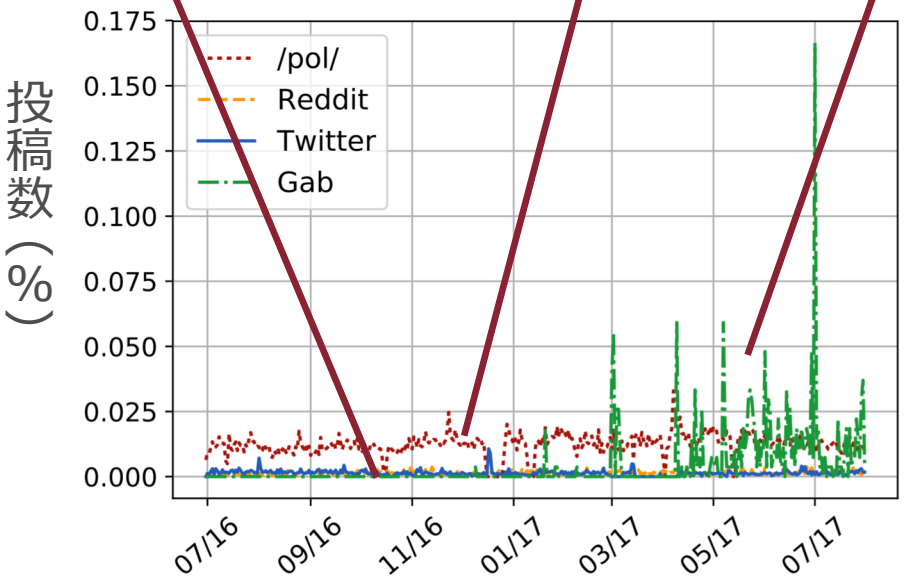
Mainstream : Twitter, Reddit
Fringe : /pol/, Gab

Mainstream で
ほとんど投稿無し

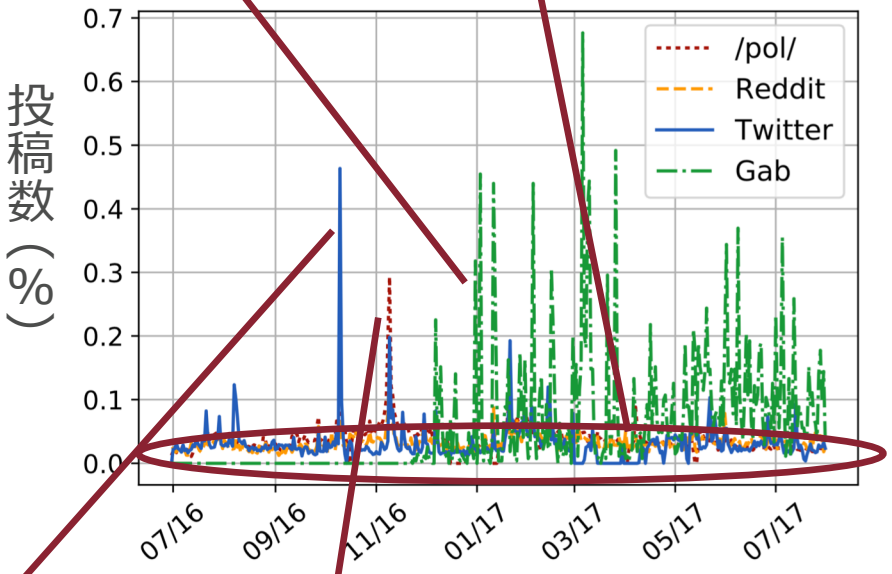
継続的な
/pol/ への投稿

2017 年
Gab の投稿増加

継続的な投稿



人種差別的ミーム



政治的ミーム

第 2 回 US 大統領選挙討論会

2016 US 大統領選挙

政治的ミームの投稿数推移は実世界の出来事と密接に関係

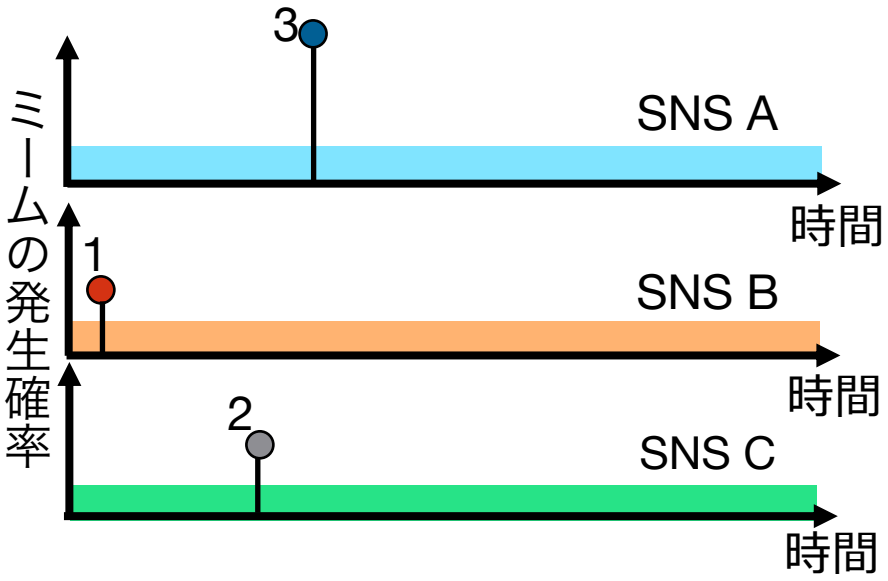
提案手法 2｜複数 SNS 間のミームの伝搬を検知

ミームの投稿が起因する SNS を調査できるモデルを利用

e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

I. 各投稿発生時にポール 1, 2, 3 を立てる



提案手法 2｜複数 SNS 間のミームの伝搬を検知

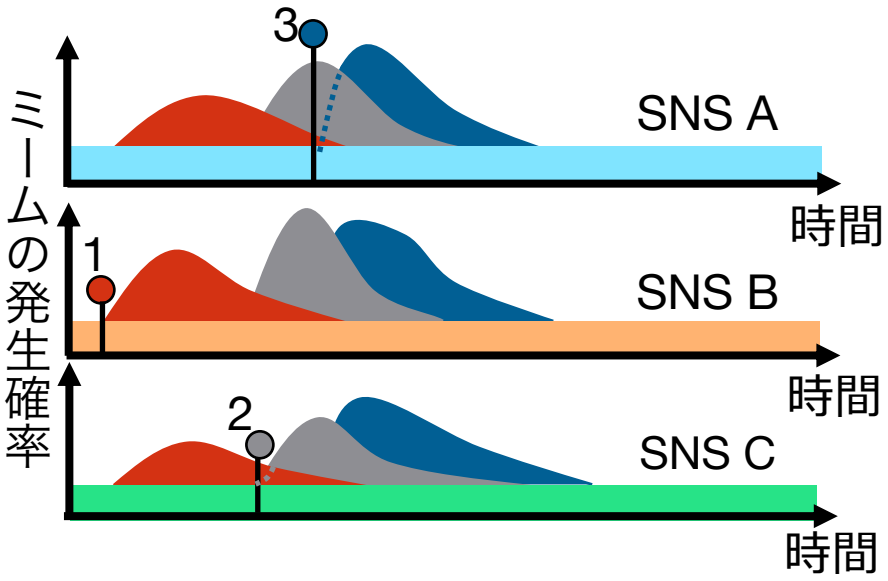
ミームの投稿が起因する SNS を調査できるモデルを利用

e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

I. 各投稿発生時にポール 1, 2, 3 を立てる

II. 各投稿発生直後, 全 SNS に波を描画



提案手法 2｜複数 SNS 間のミームの伝搬を検知

ミームの投稿が起因する SNS を調査できるモデルを利用

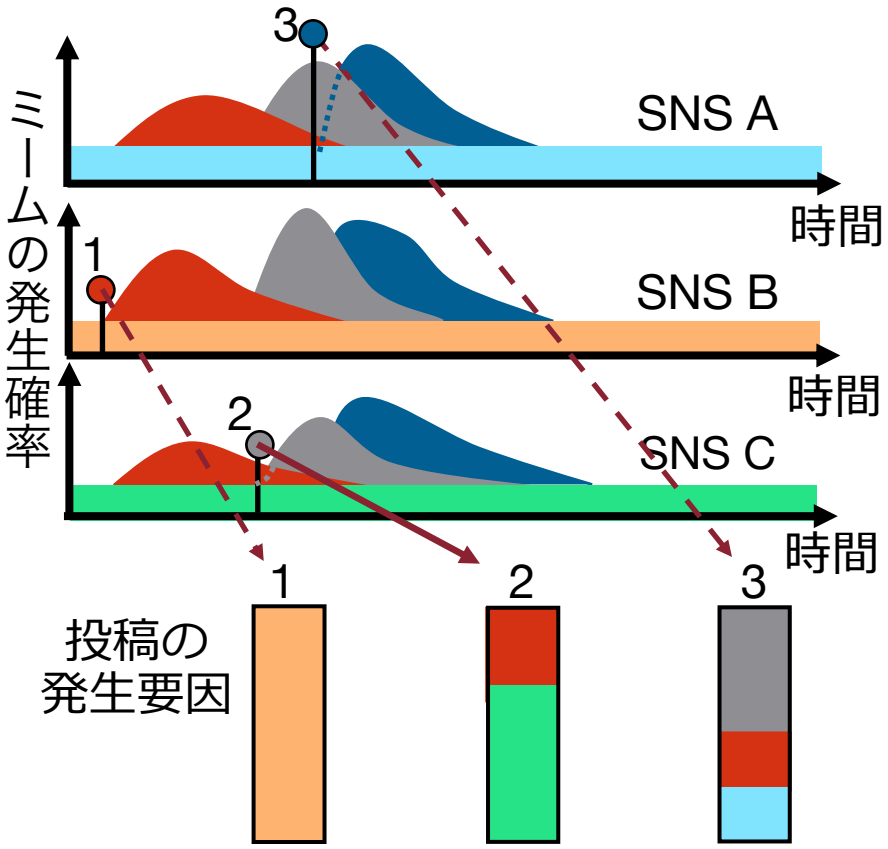
e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

I. 各投稿発生時にポール 1, 2, 3 を立てる

II. 各投稿発生直後, 全 SNS に波を描画

III. 各ポール時点での発生確率を「投稿の発生要因」として抽出



提案手法 2｜複数 SNS 間のミームの伝搬を検知

ミームの投稿が起因する SNS を調査できるモデルを利用

e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

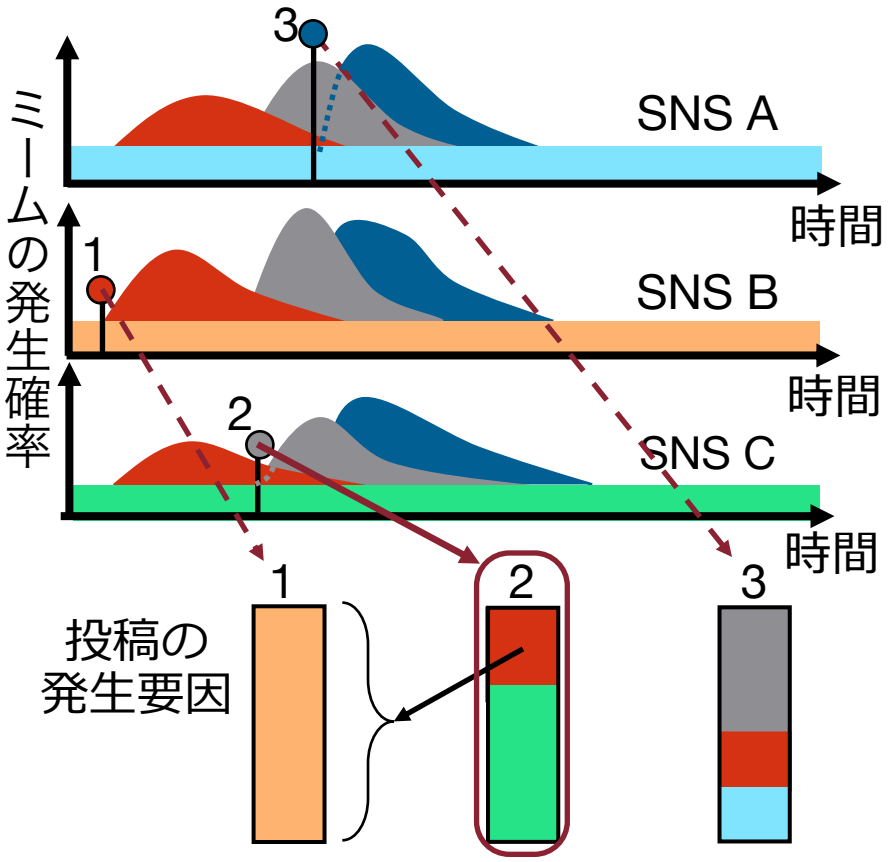
I. 各投稿発生時にポール 1, 2, 3 を立てる

II. 各投稿発生直後, 全 SNS に波を描画

III. 各ポール時点での発生確率を「投稿の発生要因」として抽出

IV. 「投稿の発生要因」の 2 を観察

- 紅色 : ポール 1 によって発生した波
- 緑色 : SNS C に起因



提案手法 2｜複数 SNS 間のミームの伝搬を検知

ミームの投稿が起因する SNS を調査できるモデルを利用

e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

I. 各投稿発生時にポール 1, 2, 3 を立てる

II. 各投稿発生直後, 全 SNS に波を描画

III. 各ポール時点での発生確率を
「投稿の発生要因」として抽出

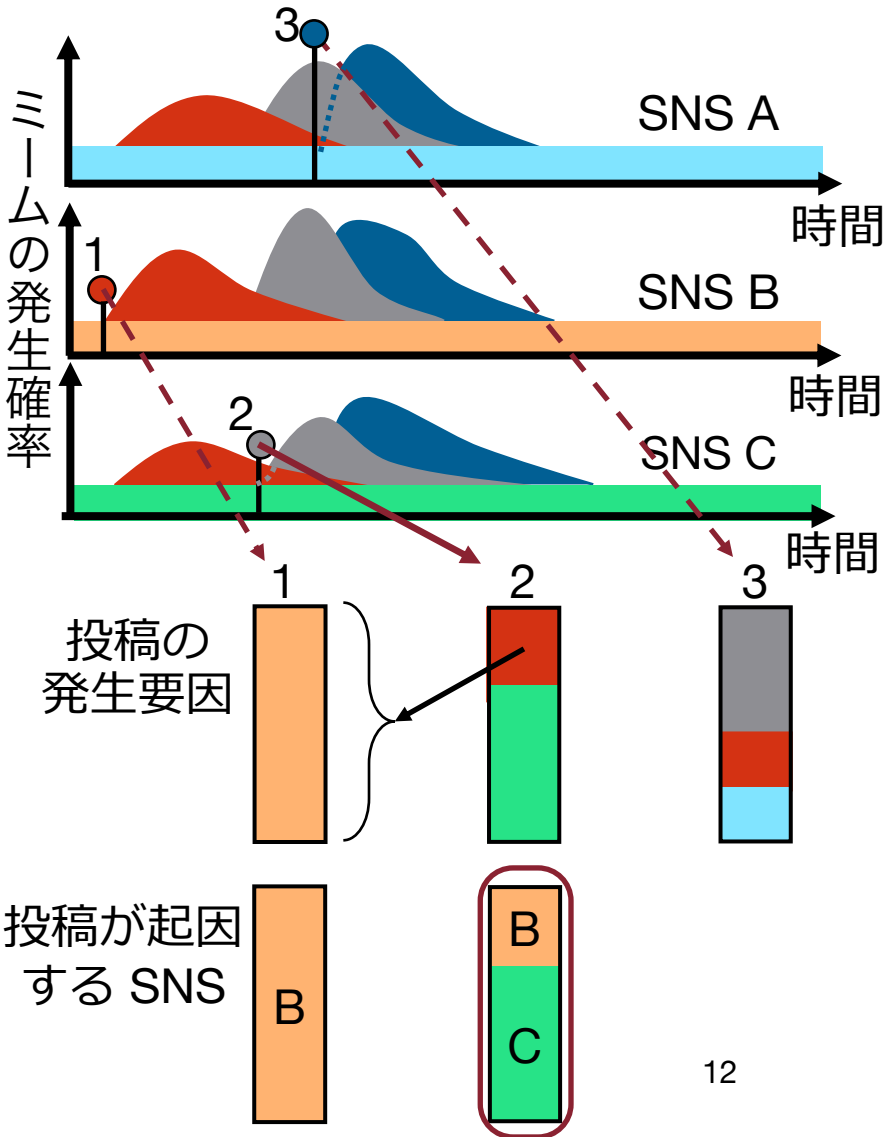
IV. 「投稿の発生要因」の 2 を観察

- 紅色 : ポール 1 によって発生した波
- 緑色 : SNS C に起因

V. 紅色を橙色に置換

- 紅色の波は SNS B (橙色) に起因

VI. **ミーム 2 の発信元 SNS が B, C と判明**



提案手法 2｜複数 SNS 間のミームの伝搬を検知

ミームの投稿が起因する SNS を調査できるモデルを利用

e.g., **ミーム 2 の発信元 SNS を特定**

- SNS B, C, A にミーム 1, 2, 3 の投稿

I. 各投稿発生時にポール 1, 2, 3 を立てる

II. 各投稿発生直後, 全 SNS に波を描画

III. 各ポール時点での発生確率を
「投稿の発生要因」として抽出

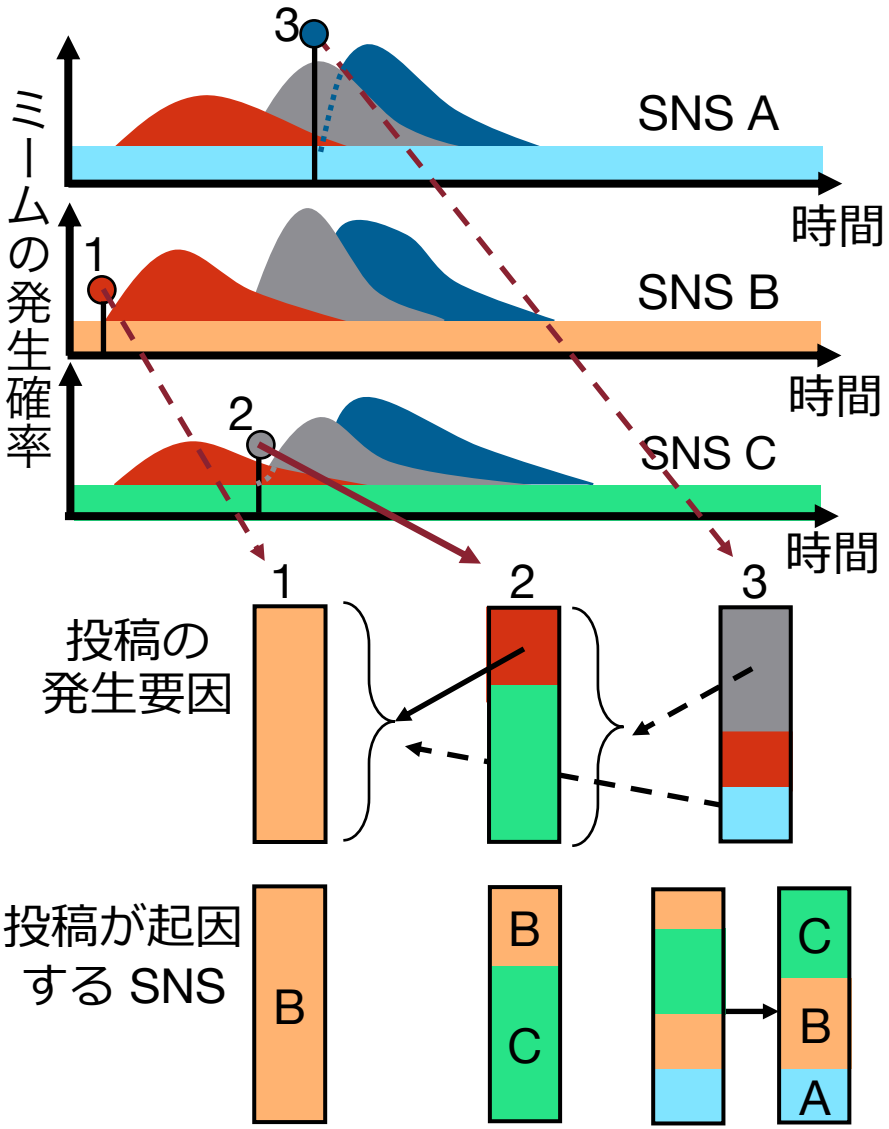
IV. 「投稿の発生要因」の 2 を観察

- 紅色 : ポール 1 によって発生した波
- 緑色 : SNS C に起因

V. 紅色を橙色に置換

- 紅色の波は SNS B (橙色) に起因

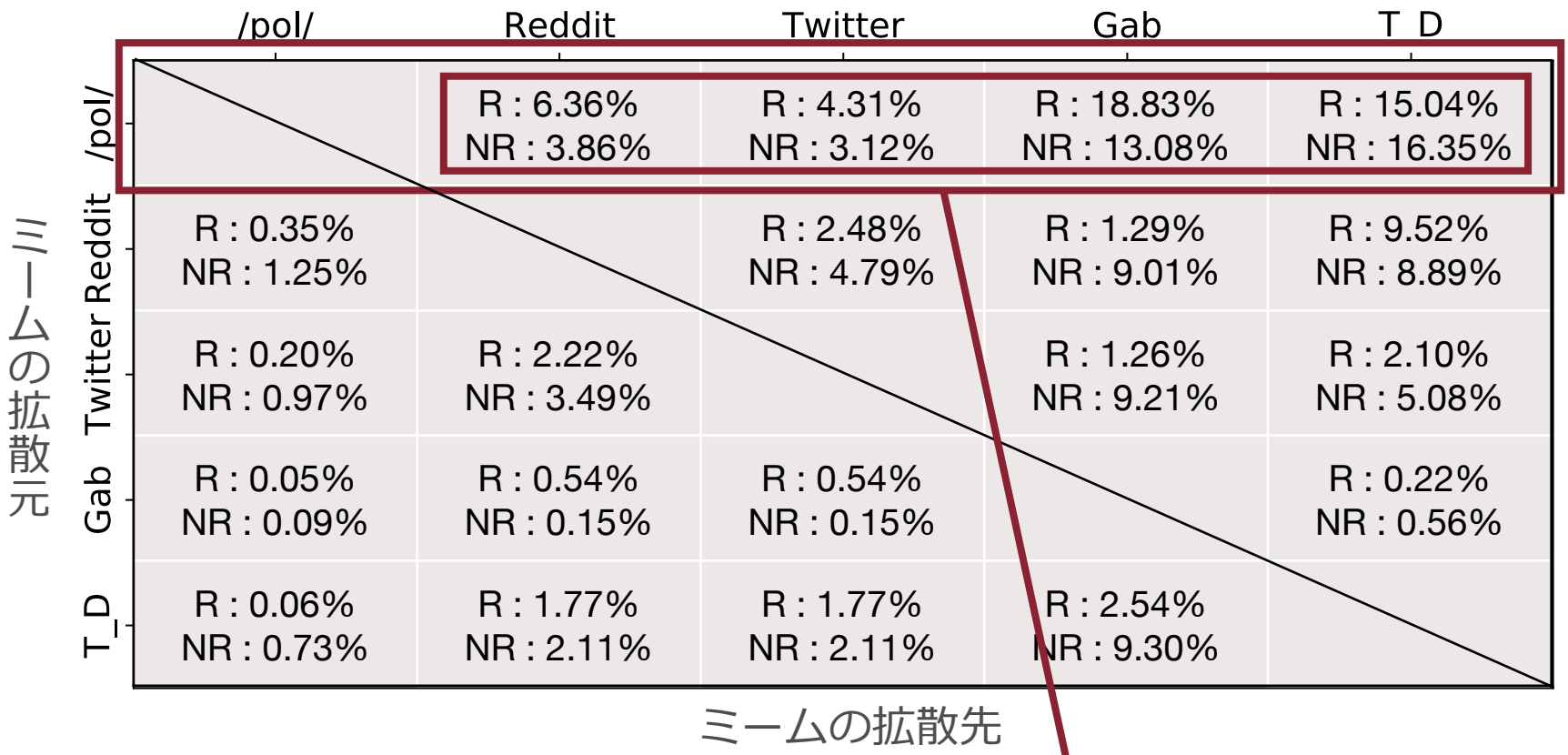
VI. **ミーム 2 の発信元 SNS が B, C と判明**



評価 2-1｜拡散された人種差別的ミームの割合

Mainstream : Twitter, Reddit
Fringe : /pol/, Gab, T_D

R： 人種差別的
NR：非人種差別的

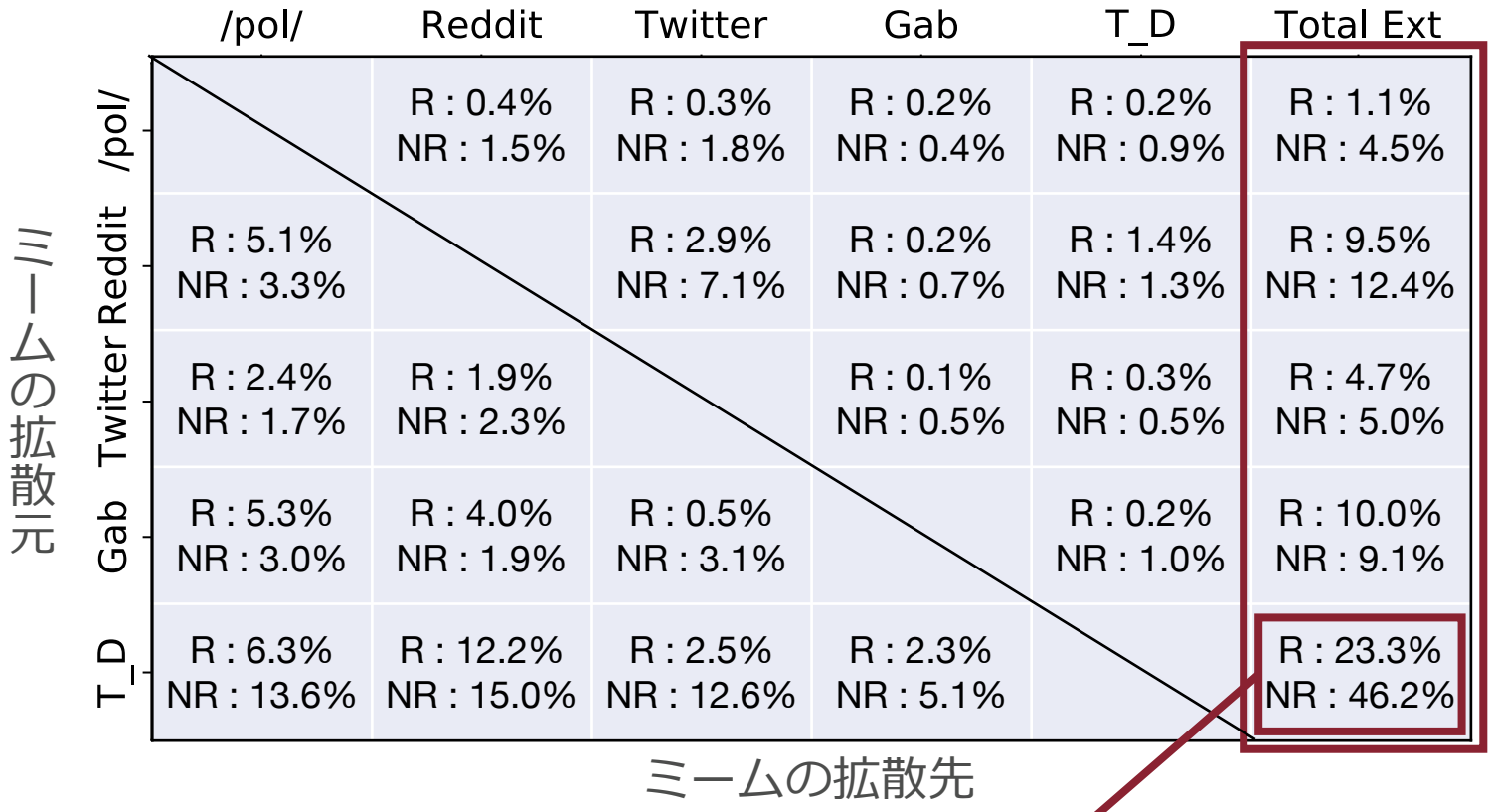


/pol/ が拡散元のとき 他の SNS への影響力が最大

評価 2-2｜人種差別的ミームが拡散される確率

Mainstream : Twitter, Reddit
Fringe : /pol/, Gab, T_D

R：人種差別的
NR：非人種差別的



T_D にミームを投稿すると他の SNS に広がる可能性が高い

まとめ

- 政治思想や信条の拡散にミームを利用
  - 攻撃的なミーム特定の需要増加
  - SNS ごとにミームの分布, 拡散の傾向の解釈が必要
- 複数 SNS のミームを意味付け
  - 人種差別的ミームが Fringe に多く投稿
  - 政治的ミームは普遍的に存在
  - 実世界の出来事は政治的ミームの投稿数に影響
- 複数 SNS 間のミームの伝搬を検知
  - /pol/ の他の SNS への影響力が最大
  - T_D にミームを投稿すると他の SNS に広がる可能性が高い

On the Origins of Memes by Means of Fringe Web Communities

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro,
Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil.

情報工学科 4 年 寺岡研究室
61619027 安森 涼

2019 年 7 月 27 日

1 背景

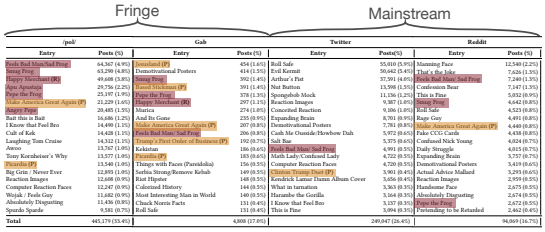
Web の台頭により考えや文化, 画像, 映像などが今までにない速度で拡散されるようになった. その中でも政治思想や心情の拡散にミーム (meme) というインターネットで流行する画像や言い回しが利用されている. その中でも人種差別的, 政治的意味をもつ攻撃的なミームの特定の需要が増加している. そこで, SNS ごとにミームの分布, 拡散の傾向の解釈が必要となっている. 本論文では画像ミームのみを対象に分析している.

2 関連研究

ミームの分析として様々な研究が行われている. 例えば, 複数の Twitter のハッシュタグからのミームの人気度調査 [1], Facebook 間の伝搬の過程で新しいミームが生まれることの証明 [2], 4chan, Reddit で口汚い言葉を用いた投稿の検知 [3] などが行われている. しかしながらこれらの事前研究では 2 点実施されていないことが 2 点存在する. まず, 1 点目として, 検証対象外の SNS データセットへの適用が不可能である点である. 具体的に言うと, Twitter に対して調査を行った手法は, Facebook などの他の SNS では利用ができないという意味である. つまり, 複数 SNS のミームを意味づけすることが必要となっている. これをハッシュ化した画像のクラスタリングを実施して解決へと導いている. 2 点目がミームの発生元となる SNS が調査されていないという点である. 具体的に言うと, Twitter で拡散されているミームがどの SNS で初めて観測されたかなどの調査はしていないと言う点である. これはつまり, 複数 SNS 間のミームの伝搬を検知する必要があると言うことで, 本論文では伝搬性を持つ事象を表現できるモデルを作成することで解決に導いている.

3 データセット

利用するデータセットとして 2 つの種類のデータセットが存在する. 1 つ目が調査対象の SNS で, 2 つ目がミームのまとめサイトである. 前者から分析対象のミームが存在する SNS, 後者からミームにメタデータを結びつけるために利用するデータを取得する. まず調査対象の SNS について, 2 項目の分類が存在する. それぞれ, Mainstream コミュニティ (以後 Mainstream と記述), Fringe コミュニティ (以後 Fringe と記述) である. それぞれ, 悪意のないミームが拡散されやすい SNS, 悪意を含むミームが拡散されやすいと考えられる SNS を示している. ここで言う悪意のあるという言葉はヘイトや人種差別のことを指す. Mainstream として Twitter と Reddit という SNS を利用する. Twitter は 140 字以内の短い記事を投稿し合うサイト, Reddit はニュース記事, 画像やテキストの投稿サイトを示す. Fringe としては The_Donald (T_D), /pol/, Gab という SNS を利用し, それぞれ, Reddit のトランプについてのチャンネル, 匿名の掲示板 4chan の政治チャンネル, ほぼ規制がない, 言論の自由を尊重するサイトを示す. ミームのまとめサイトとしては Know Your Meme (KYM) と言うものを利用する. これはミームに対するメタデータ (由来, キーワードタグ, 説明, 例, イメージギャラリー) を供給するものである.



/pol/		Gab		Twitter		Reddit	
Entry	Posts (%)	Entry	Posts (%)	Entry	Posts(%)	Entry	Posts (%)
Feels Bad Man/Sad Frog	64,367 (4.9%)	Demotivational (P)	454 (1.6%)	Roll Safe	55,010 (5.9%)	Mooning Face	12,540 (2.2%)
Smug Frog	63,290 (4.8%)	Demotivational Posters	414 (1.5%)	Evil Kermit	50,642 (5.4%)	That's the Joke	7,626 (1.3%)
Happy Merchant (R)	49,688 (3.8%)	Smug Frog	392 (1.4%)	Arthur's Fist	37,591 (4.0%)	Feels Bad Man/ Sad Frog	7,240 (1.3%)
Apu Apustaja	29,756 (2.2%)	Towel Stickmen (P)	391 (1.4%)	Nut Button	13,598 (1.5%)	Confession Bear	7,147 (1.3%)
Pepe the Frog	25,197 (1.9%)	Pepe the Frog	378 (1.3%)	Spongebob Mock	11,156 (1.2%)	This is Fine	5,032 (0.9%)
Make America Great Again (P)	21,229 (1.6%)	Happy Merchant (R)	297 (1.1%)	Reaction Images	9,387 (1.0%)	Strong Frog	4,642 (0.8%)
Angry Pepe	20,485 (1.5%)	Murica	274 (1.0%)	Conceited Reaction	9,106 (1.0%)	Roll Safe	4,523 (0.8%)
Bait (this is Bait	16,686 (1.2%)	And Its Gone	235 (0.9%)	Expanding Brain	8,701 (0.9%)	Rage Guy	4,491 (0.8%)
I Know that Feel Bro	14,499 (1.1%)	Make Amerika Great Again (P)	207 (0.8%)	Demotivational Posters	7,781 (0.8%)	Make America Great Again (P)	4,440 (0.8%)
Cult of Kek	14,428 (1.1%)	Feels Bad Man/ Sad Frog	206 (0.8%)	Cash Me Ousside/Howbow Dah	5,972 (0.6%)	Fake CGI Cards	4,438 (0.8%)
Laughing Tom Cruise	14,312 (1.1%)	Trump's First Order of Business (P)	192 (0.7%)	Salt Bae	5,375 (0.6%)	Confused Nick Young	4,024 (0.7%)
Awoo	13,767 (1.0%)	Kekistan	186 (0.6%)	Feels Bad Man/ Sad Frog	4,991 (0.5%)	Daily Struggle	4,015 (0.7%)
Tony Kornheiser's Why	13,577 (1.0%)	Picardia (P)	183 (0.6%)	Math Lady/Confused Lady	4,722 (0.5%)	Expanding Brain	3,757 (0.7%)
Picardia (P)	13,540 (1.0%)	Things with Faces (Pareidolia)	156 (0.5%)	Computer Reaction Faces	4,720 (0.5%)	Demotivational Posters	3,419 (0.6%)
Big Grin / Never Ever	12,893 (1.0%)	Serbia Strong/Remove Kebab	149 (0.5%)	Clinton Trump Duet (P)	3,901 (0.4%)	Actual Advice Mallard	3,293 (0.6%)
Reaction Images	12,688 (0.9%)	Rust Hipster	148 (0.5%)	Kendrick Lamar Duckn Album Cover	3,656 (0.4%)	Reaction Images	2,959 (0.5%)
Computer Reaction Faces	12,247 (0.9%)	Colorized History	144 (0.5%)	What in tarnation	3,363 (0.3%)	Handsome Face	2,675 (0.5%)
Wojak / Feels Guy	11,682 (0.9%)	Most Interesting Man in World	140 (0.5%)	Harambe the Gorilla	3,164 (0.3%)	Absolutely Disgusting	2,674 (0.5%)
Absolutely Disgusting	11,456 (0.8%)	Chuck Norris Facts	131 (0.4%)	I Know that Feel Bro	3,157 (0.3%)	Pepe the Frog	2,672 (0.5%)
Spurdo Sparde	9,581 (0.7%)	Roll Safe	131 (0.4%)	This is Fine	3,094 (0.3%)	Pretending to be Retarded	2,462 (0.4%)
Total	445,179 (33.4%)		4,008 (17.0%)		249,047 (26.4%)		94,869 (16.7%)

図 1 SNS ごとに投稿されるミームの割合

4 提案手法 1 | 複数 SNS のミームの意味付け, 評価

ある要素の出力が次の要素の入力となるような処理を繰り返すことで複数 SNS のミームに意味付けを実施する. まずミームの一部を似た画像が近い値を保持するという特徴をもつ 64bit ハッシュ値を用いて, ハッシュ化をする. 続いて, ハミング距離準拠のアルゴリズムを利用してクラスタリングを行い, クラスタごとに Medoid と言うクラスタ内の自身以外の画像との距離の総和が最小となる, クラスタを代表する画像を決定する. 続いて, ハッシュ化した KYM データを利用し, クラスタとメタデータを結びつけ, その結果を他のデータセットへ適用する. この処理の入力が画像のため任意の SNS に対応可能であり, 前述の検証対象外の SNS データセットへの適用が不可能という問題を解決している.

4.1 SNS ごとに投稿されるミームの割合

前述の処理を行った後に出力された結果を元に分析を行い, 出力された結果が図 1 である. この図において, 左 2 つ (/pol/, Gab) が Fringe を, 右 2 つ (Twitter, Reddit) が Mainstream を示している. これを見ると人種差別的ミーム (図中赤) が Fringe で上位であるのに対して Mainstream では悪意のないミームが上位を占めている. ここで述べた人種差別的というのは'racist', 'antisemitism'(反ユダヤ主義) などのタグを持つものである. 一方で政治的なミームは Mainstream, Fringe 問わずどこにでも存在していることがわかる.

4.2 ミームの投稿数の推移

人種差別的ミームおよび政治的ミームの投稿数の推移はそれぞれ図 2, 図 3 で表される. 共に 2017 年に Gab の投稿が増加していることが観察できる. これは Gab が 2016 年後期にサービスを開始し, 2017 年に徐々に普及してきたことが原因として考えられる. 続いて, 人種差別的ミームについて観察すると, Mainstream ではほとんど投稿がないのに対し, /pol/(Fringe) では継続的に投稿があることがわかる. 一方で政治的ミームに関しては Mainstream, Fringe に関係なく全てのコミュニティで継続的な投稿があることがわかる. これは 4.1 節での評価の裏付けとなっている. また, 政治的ミームについては第 2 回 US 大統領選挙討論会や 2016 US 大統領選挙において投稿数が伸びていることから政治的ミームの投稿数推移は実世界の出来事と密接に関係していることがわかる.

5 提案手法 2 | 複数 SNS 間のミームの伝搬を検知

複数 SNS 間のミームの伝搬の検知にはミームの投稿が起因する SNS を調査できるモデルを作成し利用することで解決に導いてい

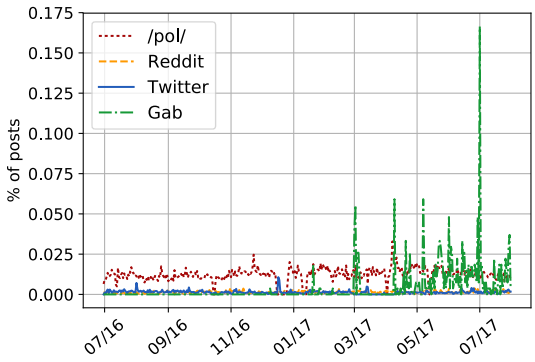


図 2　人種差別的ミームの投稿数の推移

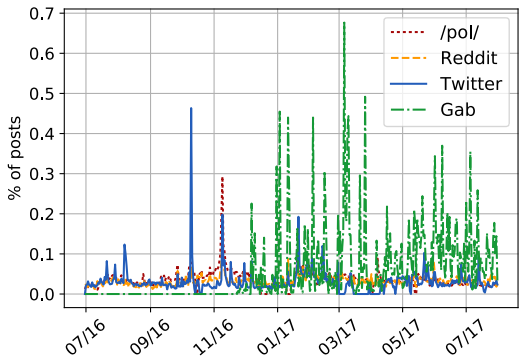


図 3　政治的ミームの投稿数の推移

る. 本項では説明のため SNS B, C, A に順にミーム 1, 2, 3 の投稿があった場合にミーム 2 の発信元 SNS を特定することを考える. 図 4 を元に説明する. まず各投稿発生時にポール 1, 2, 3 を立て, 各投稿発生直後全ての SNS に波を描画する. そして各ポール時点での発生確率を「投稿の発生要因」として抽出し, 「投稿の発生要因」の 2 について観察する. 観察を行うと「投稿の発生要因」の 2 の紅色で表されている部分はポール 1 によって発生した波に起因する部分で, 緑色で表されている部分は SNS C に起因していることがわかる. 紅色で表される部分は SNS B に 100% 起因している部分であるから, 紅色を橙色に置換した結果を投稿が起因する SNS として表示する. これによりミーム 2 の発信元 SNS が B, C と判明する. 同様の手順でミーム 3 についても投稿が起因する SNS を求めることが可能である.

5.1　拡散された人種差別的ミームの割合

前述の手順で処理を行った後に出力された結果を元に拡散された人種差別的ミームを求めたものが図 5 である. 図の縦はミームの拡散元を, 横はミームの拡散先を, 図は拡散されたミームの割合を示している. 図の R と NR は人種差別的, 非人種差別的なミームを示している. この図を観察すると, /pol/が拡散元の時他の SNS への影響力が高いことがわかる. つまり他のコミュニティに広まったミームに/pol/発祥のものが多いことがわかる.

5.2　人種差別的ミームが拡散される確率

図 6 は 5.1 の図について送信元の SNS で正規化したものである. この図より T_D が送信元の時に拡散される確率が高いことから, T_D にミームを投稿すると他の SNS に広がる可能性が高いことがわかる. 一方で, 影響力が最大であった/pol/は拡散される確率という面では最悪で, 投稿のうち一流のものしか残らないことがわかる. 政治的ミームについても本論文で言及があったが少しの変化しかなかったため割愛する.

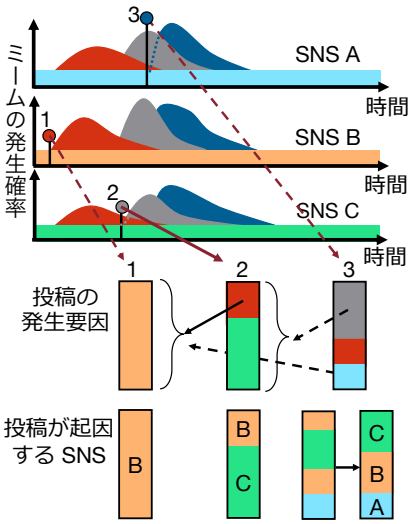


図 4　政治的ミームの投稿数の推移

	R : 6.36% NR : 3.86%	R : 4.31% NR : 3.12%	R : 18.83% NR : 13.08%	R : 15.04% NR : 16.35%
R : 0.35% NR : 1.25%		R : 2.48% NR : 4.79%	R : 1.29% NR : 9.01%	R : 9.52% NR : 8.89%
R : 0.20% NR : 0.97%	R : 2.22% NR : 3.49%		R : 1.26% NR : 9.21%	R : 2.10% NR : 5.08%
R : 0.05% NR : 0.09%	R : 0.54% NR : 0.15%	R : 0.54% NR : 0.15%		R : 0.22% NR : 0.56%
R : 0.06% NR : 0.73%	R : 1.77% NR : 2.11%	R : 1.77% NR : 2.11%	R : 2.54% NR : 9.30%	

図 5　拡散された人種差別的ミームの割合

	R : 0.4% NR : 1.5%	R : 0.3% NR : 1.8%	R : 0.2% NR : 0.4%	R : 0.2% NR : 0.9%	R : 1.1% NR : 4.5%
R : 5.1% NR : 3.3%		R : 2.9% NR : 7.1%	R : 0.2% NR : 0.7%	R : 1.4% NR : 1.3%	R : 9.5% NR : 12.4%
R : 2.4% NR : 1.7%	R : 1.9% NR : 2.3%		R : 0.1% NR : 0.5%	R : 0.3% NR : 0.5%	R : 4.7% NR : 5.0%
R : 5.3% NR : 3.0%	R : 4.0% NR : 1.9%	R : 0.5% NR : 3.1%		R : 0.2% NR : 1.0%	R : 10.0% NR : 9.1%
R : 6.3% NR : 13.6%	R : 12.2% NR : 15.0%	R : 2.5% NR : 12.6%	R : 2.3% NR : 5.1%		R : 23.3% NR : 46.2%

図 6　人種差別的ミームが拡散される確率

6　結論

政治思想や信条の拡散にミームが利用され, その中でも人種差別的, 政治的な意味を含む攻撃的なミームの特定の需要が増えている. そこで, SNS ごとにミームの分布や拡散の傾向の解釈が必要であると考えられている. 本論文では複数 SNS のミームの意味付けおよび複数 SNS 間のミームの伝搬を検知している. 結果として人種差別的ミームが Fringe に多く投稿されている点, 政治的ミームが普遍的に存在している点, 実世界の出来事は政治的ミームの投稿数に影響する点, /pol/の他の SNS への影響力が最大である点, T_D にミームを投稿すると他の SNS に広がる可能性が高い点などが判明した.

7　参考文献

[1] L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting Successful Memes Using Network and Community Structure. In ICWSM, 2014.

[2] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng. Information Evolution in Social Networks. In WSDM, 2016.

[3] E. Chandrasekharan, et al., The bag of communities: Identifying abusive behavior online with preexisting Internet data. In CHI, 2017.

On the Origins of Memes by Means of Fringe Web Communities

Savvas Zannettou[★], Tristan Caulfield[‡], Jeremy Blackburn[†], Emiliano De Cristofaro[‡], Michael Sirivianos[★], Gianluca Stringhini[◇], and Guillermo Suarez-Tangil[+]

[★]Cyprus University of Technology, [‡]University College London,
[†]University of Alabama at Birmingham, [◇]Boston University, [+]King's College London

sa.zannettou@edu.cut.ac.cy, {t.caulfield,e.decristofaro}@ucl.ac.uk, blackburn@uab.edu, michael.sirivianos@cut.ac.cy, gian@bu.edu, guillermo.suarez-tangil@kcl.ac.uk

ABSTRACT

Internet memes are increasingly used to sway and manipulate public opinion. This prompts the need to study their propagation, evolution, and influence across the Web. In this paper, we detect and measure the propagation of memes across multiple Web communities, using a processing pipeline based on perceptual hashing and clustering techniques, and a dataset of 160M images from 2.6B posts gathered from Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab, over the course of 13 months. We group the images posted on fringe Web communities (/pol/, Gab, and The_Donald subreddit) into clusters, annotate them using meme metadata obtained from Know Your Meme, and also map images from mainstream communities (Twitter and Reddit) to the clusters.

Our analysis provides an assessment of the popularity and diversity of memes in the context of each community, showing, e.g., that racist memes are extremely common in fringe Web communities. We also find a substantial number of politics-related memes on both mainstream and fringe Web communities, supporting media reports that memes might be used to enhance or harm politicians. Finally, we use Hawkes processes to model the interplay between Web communities and quantify their reciprocal influence, finding that /pol/ substantially influences the meme ecosystem with the number of memes it produces, while The_Donald has a higher success rate in pushing them to other communities.

CCS CONCEPTS

• **General and reference** → **General conference proceedings**; **Measurement**; • **Mathematics of computing** → **Multivariate statistics**; • **Networks** → **Social media networks**; **Online social networks**; • **Computing methodologies** → *Neural networks*;

KEYWORDS

Memes, Influence, Twitter, Reddit, 4chan, Gab

IMC '18, October 31-November 2, 2018, Boston, MA, USA

ACM ISBN 978-1-4503-5619-0/18/10…$15.00
https://doi.org/10.1145/3278532.3278550

ACM Reference Format:
Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In Proceedings of IMC '18. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3278532.3278550

1 INTRODUCTION

The Web has become one of the most impactful vehicles for the propagation of ideas and culture. Images, videos, and slogans are created and shared online at an unprecedented pace. Some of these, commonly referred to as *memes*, become viral, evolve, and eventually enter popular culture. The term "meme" was first coined by Richard Dawkins [11], who framed them as cultural analogues to genes, as they too self-replicate, mutate, and respond to selective pressures [17]. Numerous memes have become integral part of Internet culture, with well-known examples including the Trollface [49], Bad Luck Brian [25], and Rickroll [44].

While most memes are generally ironic in nature, used with no bad intentions, others have assumed negative and/or hateful connotations, including outright racist and aggressive undertones [71]. These memes, often generated by fringe communities, are being "weaponized" and even becoming part of political and ideological propaganda [61]. For example, memes were adopted by candidates during the 2016 US Presidential Elections as part of their iconography [18]; in October 2015, then-candidate Donald Trump retweeted an image depicting him as Pepe The Frog, a controversial character considered a hate symbol [53]. In this context, polarized communities within 4chan and Reddit have been working hard to create new memes and make them go viral, aiming to increase the visibility of their ideas—a phenomenon known as "attention hacking" [59].

Despite their increasingly relevant role, we have very little measurements and computational tools to understand the origins and the influence of memes. The online information ecosystem is very complex; social networks do not operate in a vacuum but rather influence each other as to how information spreads [74]. However, previous work (see Sec. 6) has mostly focused on social networks in an isolated manner.

In this paper, we aim to bridge these gaps by identifying and addressing a few research questions, which are oriented towards fringe Web communities: 1) How can we characterize memes, and how do they evolve and propagate? 2) Can we track meme propagation across multiple communities and measure their influence?

3) How can we study variants of the same meme? 4) Can we characterize Web communities through the lens of memes?

Our work focuses on four Web communities: Twitter, Reddit, Gab, and 4chan's Politically Incorrect board (/pol/), because of their impact on the information ecosystem [74] and anecdotal evidence of them disseminating weaponized memes [65]. We design a processing pipeline and use it over 160M images posted between July 2016 and July 2017. Our pipeline relies on perceptual hashing (pHash) and clustering techniques; the former extracts representative feature vectors from the images encapsulating their visual peculiarities, while the latter allow us to detect groups of images that are part of the same meme. We design and implement a custom distance metric, based on both pHash and meme metadata, obtained from Know Your Meme (KYM), and use it to understand the interplay between the different memes. Finally, using Hawkes processes, we quantify the reciprocal influence of each Web community with respect to the dissemination of image-based memes.

Findings. Some of our findings (among others) include:

(1) Our influence estimation analysis reveals that /pol/ and The_Donald are influential actors in the meme ecosystem, despite their modest size. We find that /pol/ substantially influences the meme ecosystem by posting a large number of memes, while The_Donald is the most *efficient* community in pushing memes to both fringe and mainstream Web communities.
(2) Communities within 4chan, Reddit, and Gab use memes to share hateful and racist content. For instance, among the most popular cluster of memes, we find variants of the antisemitic "Happy Merchant" meme [32] and the controversial Pepe the Frog [40].
(3) Our custom distance metric effectively reveals the phylogenetic relationships of clusters of images. This is evident from the graph that shows the clusters obtained from /pol/, Reddit's The_Donald subreddit, and Gab available for exploration at [1].

Contributions. First, we develop a robust processing pipeline for detecting and tracking memes across multiple Web communities. Based on pHash and clustering algorithms, it supports large-scale measurements of meme ecosystems. Second, we introduce a custom distance metric, geared to highlight hidden correlations between memes and better understand the interplay and overlap between them. Third, we provide a characterization of multiple Web communities (Twitter, Reddit, Gab, and /pol/) with respect to the memes they share, and an analysis of their reciprocal influence using the Hawkes Processes statistical model. Finally, we release our processing pipeline and datasets (available at [3]), hence supportting further measurements in this space.

2 METHODOLOGY

In this section, we present our methodology for measuring the propagation of memes across Web communities.

2.1 Overview

Memes are high-level concepts or ideas that spread within a culture [11]. In Internet vernacular, a *meme* usually refers to variants of a particular image, video, cliché, etc. that share a common theme and are disseminated by a large number of users. In this paper, we focus on their most common incarnation: *static images*.

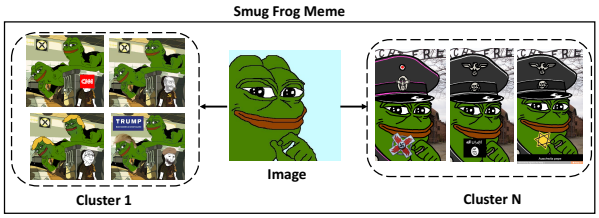


Figure 1: An example of a meme (Smug Frog) that provides an intuition of what an image, a cluster, and a meme is.

To gain an understanding of how memes propagate across the Web, with a particular focus on discovering the communities that are most influential in spreading them, our intuition is to build *clusters* of visually similar images, allowing us to track variants of a meme. We then group clusters that belong to the same meme to study and track the meme itself. In Fig. 1, we provide a visual representation of the Smug Frog meme [47], which includes many variants of the same image and several clusters of variants. Cluster 1 has variants from a Jurassic Park scene, where one of the characters is hiding from two velociraptors behind a kitchen counter: the frogs are stylized to look similar to velociraptors, and the character hiding varies to express a particular message. For example, in the image in the top right corner, the two frogs are searching for an antisemitic caricature of a Jew. Cluster N shows variants of the smug frog wearing a Nazi officer military cap with the infamous "Arbeit macht frei" in the background. Overall, these clusters represent the branching nature of memes: as a new variant of a meme becomes prevalent, it often branches into its own sub-meme, potentially incorporating imagery from other memes.

2.2 Processing Pipeline

We now introduce our processing pipeline; see Fig. 2. As discussed in Sec. 2.1, our methodology aims at identifying clusters of similar images and assign them to higher level groups, which are the actual memes. Note that the proposed pipeline is not limited to image macros and can be used to identify any image. We first discuss the types of data sources needed for our approach, i.e., meme annotation sites and Web communities that post memes (dotted rounded rectangles in the figure). Then, we describe each of the operations performed by our pipeline (Steps 1-7, see regular rectangles).

Data Sources. Our pipeline uses two types of data sources: 1) sites providing meme annotation and 2) Web communities that disseminate memes. In this paper, we use Know Your Meme for the former, and Twitter, Reddit, /pol/, and Gab for the latter (see Sec. 3). However, our methodology supports any annotation site and any Web community, and this is why we add the "*Generic*" sites/communities notation in Fig. 2.

pHash Extraction (Step 1). We use the Perceptual Hashing (pHash) algorithm [60] to calculate a fingerprint of each image in such a way that any two images that look similar to the human eye map to a "similar" hash value. pHash generates a feature vector of 64 elements that describe an image, computed from the Discrete Cosine Transform among the different frequency domains of the image.

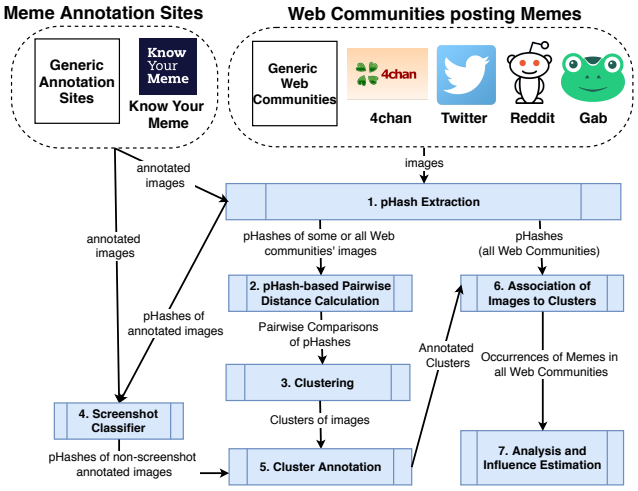


Figure 2: High-level overview of our processing pipeline.

Thus, visually similar images have minor differences in their vectors. For example, the string representation of the phashes obtained from the images in cluster N (see Fig. 1) are 55352b0b8d8b5b53, 55952b0bb58b5353, and 55952b2b9da58a53, respectively. The algorithm is also robust against changes in the images, e.g., signal processing operations and direct manipulation [75], and effectively reduces the dimensionality of the raw images.

Clustering via pairwise distance calculation (Steps 2-3). Next, we cluster images from one or more Web Communities using the pHash values. We perform a pairwise comparison of all the pHashes using Hamming distance (Step 2). To support large numbers of images, we implement a highly parallelizable system on top of TensorFlow [5], which uses multiple GPUs to enhance performance. Images are clustered using a density-based algorithm (Step 3). Our current implementation uses DBSCAN [14], mainly because it can discover clusters of arbitrary shape and performs well over large, noisy datasets. Nonetheless, our architecture can be easily tweaked to support any clustering algorithm and distance metric.

We also perform an analysis of the clustering performance and the rationale for selecting the clustering threshold. We refer to Appendix A for more details.

Screenshots Removal (Step 4). Meme annotation sites like KYM often include, in their image galleries, screenshots of social network posts that are not variants of a meme but just comments about it. Hence, we discard social-network screenshots from the annotation sites data sources using a deep learning classifier. Due to space limitations, we refer to Appendix C of the extended version of the paper [73] for details on the classifier.

Cluster Annotation (Steps 5). Clustering annotation uses the *medoid* of each cluster, i.e., the element with the minimum square average distance from all images in the cluster. In other words, the medoid is the image that best represents the cluster. The clusters' medoids are compared with all images from meme annotation sites, by calculating the Hamming distance between each pair of pHash vectors. We consider that an image matches a cluster if the distance is less than or equal to a threshold θ, which we set to 8, as it allows us to capture the diversity of images that are part of the same meme while maintaining a low number of false positives.

As the annotation process considers all the images of a KYM entry's image gallery, it is likely we will get multiple annotations for a single cluster. To find the representative KYM entry for each cluster, we select the one with the largest proportion of matches of KYM images with the cluster medoid. In case of ties, we select the one with the minimum average Hamming distance.

As KYM is based on community contributions it is unclear how good our annotations are. To evaluate KYM entries and our cluster annotations, three authors of this paper assessed 200 annotated clusters and 162 KYM entries. We find that only a 1.85% of the assessed KYM entries were regarded as "bad" or not sufficient. When it comes to the clustering annotation, we note that the three annotators had substantial agreement (Fleis agreement score equal to 0.67) and that the clustering accuracy, after majority agreement, of the assessed clusters is 89% . We refer to Appendix B for details about the annotation process and results.

Association of images to memes (Step 6). To associate images posted on Web communities (e.g., Twitter, Reddit, etc.) to memes, we compare them with the clusters' medoids, using the same threshold θ. This is conceptually similar to Step 5, but uses images from Web communities instead of images from annotation sites. This lets us identify memes posted in generic Web communities and collect relevant metadata from the posts (e.g., the timestamp of a tweet). Note that we track the propagation of memes in generic Web communities (e.g., Twitter) using a *seed* of memes obtained by clustering images from other (fringe) Web communities. More specifically, our seeds will be memes generated on three fringe Web communities (/pol/, The_Donald subreddit, Gab); nonetheless, our methodology can be applied to any community.

Analysis and Influence Estimation (Step 7). We analyze all relevant clusters and the occurrences of memes, aiming to assess: 1) their popularity and diversity in each community; 2) their temporal evolution; 3) how communities influence each other with respect to meme dissemination (see Sec. 4 and 5).

2.3 Distance Metric

To better understand the interplay and connections between the clusters, we introduce a custom distance metric, which relies on both the visual peculiarities of the images (via pHash) and data available from annotation sites. The distance metric supports one of two modes: 1) one for when both clusters are annotated (*full-mode*), and 2) another for when one or none of the clusters is annotated (*partial-mode*).

Definition. Let c be a cluster of images and F a set of features extracted from the clusters. The custom distance metric between cluster c_i and c_j is defined as:

$$\text{distance}(c_i, c_j) = 1 - \sum_{\mathsf{f} \in \mathsf{F}} w_\mathsf{f} \times \mathsf{r}_\mathsf{f}(c_i, c_j) \quad (1)$$

where $\mathsf{r}_\mathsf{f}(c_i, c_j)$ denotes the similarity between the features of type $\mathsf{f} \in \mathsf{F}$ of cluster c_i and c_j, and w_f is a weight that represents the relevance of each feature. Note that $\sum_\mathsf{f} w_\mathsf{f} = 1$ and $\mathsf{r}_f(c_i, c_j) = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Thus, $\text{distance}(c_i, c_j)$ is a number between 0 and 1.

Features. We consider four different features for $r_{f \in F}$, specifically, $F = \{perceptual, meme, people, culture\}$; see below.

$r_{perceptual}$: this feature is the similarity between two clusters from a perceptual viewpoint. Let h be a pHash vector for an image m in cluster c, where m is the medoid of the cluster, and d_{ij} the Hamming distance between vectors h_i and h_j (see in Step 5). We compute d_{ij} from c_i and c_j as follows. First, we obtain obtain the medoid m_i from cluster c_i. Subsequently, we obtain h_i=pHash(m_i). Finally, we compute d_{ij}=Hamming(h_i, h_j). We simplify notation and use d instead of d_{ij} to denote the distance between two medoid images and refer to this distance as the Hamming *score*.

We define the perceptual similarity between two clusters as an exponential decay function over the Hamming score d:

$$r_{perceptual}(d) = 1 - \frac{d}{\tau \times e^{\max/\tau}} \quad (2)$$

where max represents the maximum pHash distance between two images and τ is a constant parameter, or *smoother*, that controls how fast the exponential function decays for all values of d (recall that $\{d \in \mathbb{R} \mid 0 \leq d \leq \max\}$). Note that max is bound to the precision given by the pHash algorithm. Recall that each pHash has a size of $|d|$=64, hence max=64. Intuitively, when $\tau << 64$, $r_{perceptual}$ is a high value only with perceptually indistinguishable images, e.g., for τ=1, two images with d=0 have a similarity $r_{perceptual}$=1.0. With the same τ, the similarity drops to 0.4 when d=1. By contrast, when τ is close to 64, $r_{perceptual}$ decays almost linearly. For example, for τ=64, $r_{perceptual}(d=0)$=1.0 and $r_{perceptual}(d=1)$=0.98. As mentioned above, we observe that pairs of images with scores between d=0 and d=8 are usually part of the same variant (see Step 5 in Sec. 2.2). In our implementation, we set τ=25 as $r_{perceptual}$ returns high values up to d=8, and rapidly decays thereafter.

r_{meme}, $r_{culture}$, and r_{people}: the annotation process (Step 5) provides contextualized information about the cluster medoid, including the name (i.e., the main identifier) given to a meme, the associated culture (i.e., high-level group of meme), and people that are included in a meme. (Note that we use all the annotations for each category and not only the representative one, see Step 5.) Therefore, we model a different similarity for each of the these categories, by looking at the overlap of all the annotations among the medoids of both clusters (m_i, m_j, for c_i and c_j, respectively). Specifically, for each category, we calculate the Jaccard index between the annotations of both medoids, for memes, cultures, and people, thus acquiring r_{meme}, $r_{culture}$, r_{people}, respectively.

Modes. Our distance metric measures how similar two clusters are. If both clusters are annotated, we operate in "full-mode," and in "partial-mode" otherwise. For each mode, we use different weights for the features in Eq. 1, which we set empirically as we lack the ground-truth data needed to automate the computation of the optimal set of thresholds.

Full-mode. In full-mode, we set weights as follows. 1) The features from the perceptual and meme categories should have higher relevance than people and culture, as they are intrinsically related to the definition of meme (see Sec. 2.1). The last two are non-discriminant features, yet are informative and should contribute to the metric. Also, 2) r_{meme} should not outweigh $r_{perceptual}$ because of the relevance that visual similarities have on the different variants of a meme. Likewise, $r_{perceptual}$ should not dominate over r_{meme} because of the branching nature of the memes. Thus, we want these two categories to play an equally important weight. Therefore, we choose $w_{perceptual}$=0.4, w_{meme}=0.4, w_{people}=0.1, $w_{culture}$=0.1.

This means that when two clusters belong to the same meme and their medoids are perceptually similar, the distance between the clusters will be small. In fact, it will be at most $0.2 = 1 - (0.4 + 0.4)$ if people and culture do not match, and 0.0 if they also match. Note that our metric also assigns small distance values for the following two cases: 1) when two clusters are part of the same meme variant, and 2) when two clusters use the same image for different memes.

Partial-mode. In this mode, we associate unannotated images with any of the known clusters. This is a critical component of our analysis (Step 6), allowing us to study images from generic Web communities where annotations are unavailable. In this case, we rely entirely on the perceptual features. We once again use Eq. 1, but simply set all weights to 0, except for $w_{perceptual}$ (which is set to 1). That is, we compare the image we want to test with the medoid of the cluster and we apply Eq. 2 as described above.

Platform	#Posts	#Posts with Images	#Images	#Unique pHashes
Twitter	1,469,582,378	242,723,732	114,459,736	74,234,065
Reddit	1,081,701,536	62,321,628	40,523,275	30,441,325
/pol/	48,725,043	13,190,390	4,325,648	3,626,184
Gab	12,395,575	955,440	235,222	193,783
KYM	15,584	15,584	706,940	597,060

Table 1: Overview of our datasets.

3 DATASETS

3.1 Web Communities

As mentioned earlier, our data sources are Web communities that post memes and meme annotation sites. For the former, we focus on four communities: Twitter, Reddit, Gab, and 4chan (more precisely, 4chan's Politically Incorrect board, /pol/). This provides a mix of mainstream social networks (Twitter and Reddit) as well as fringe communities that are often associated with the alt-right and have an impact on the information ecosystem (Gab and /pol/) [74].

There are several other platforms playing important roles in spreading memes, however, many are "closed" (e.g., Facebook) or do not involve memes based on static images (e.g., YouTube, Giphy). In future work, we plan to extend our measurements to communities like Instagram and Tumblr, as well as to GIF and video memes. Nonetheless, we believe our data sources already allow us to elicit comprehensive insights into the meme ecosystem.

Table 1 reports the number of posts and images processed for each community. Note that the number of images is lower than the number of posts with images because of duplicate image URLs and because some images get deleted. Next, we discuss each dataset.

Twitter. Twitter is a mainstream microblogging platform, allowing users to broadcast 280-character messages (tweets) to their followers. Our Twitter dataset is based on tweets made available via the 1% Streaming API, between July 1, 2016 and July 31, 2017. In total, we parse 1.4B tweets: 242M of them have at least one image. We

extract all the images, ultimately collecting 114M images yielding 74M unique pHashes.

Reddit. Reddit is a news aggregator: users create submissions by posting a URL and others can reply in a structured way. It is divided into multiple sub-communities called subreddits, each with its own topic and moderation policy. Content popularity and ranking are determined via a voting system based on the up- and down-votes users cast. We gather images from Reddit using publicly available data from Pushshift [62]. We parse all submissions and comments[1] between July 1, 2016 and July, 31 2017, and extract 62M posts that contain at least one image. We then download 40M images producing 30M unique pHashes.

4chan. 4chan is an anonymous image board; users create new threads by posting an image with some text, which others can reply to. It has two characteristic features: anonymity and ephemerality. By default, user identities are concealed, and all threads are deleted after one week. Overall, 4chan is known for its extremely lax moderation and the high degree of hate and racism, especially on boards like /pol/ [20]. We obtain all threads posted on /pol/, between July 1, 2016 and July 31, 2017, using the same methodology of [20]. Since all threads (and images) are removed after a week, we use a public archive service called 4plebs [4] to collect 4.3M images, thus yielding 3.6M unique pHashes.

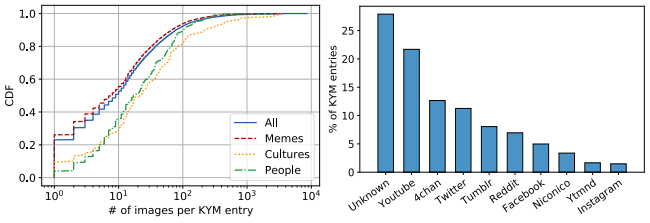
Gab. Gab is a social network launched in August 2016 as a "champion" of free speech, providing "shelter" to users banned from other platforms. It combines social networking features from Twitter (broadcast of 300-character messages) and Reddit (content is ranked according to up- and down-votes). It also has extremely lax moderation as it allows everything except illegal pornography, terrorist propaganda, and doxing [66]. Overall, Gab attracts alt-right users, conspiracy theorists, and trolls, and high volumes of hate speech [72]. We collect 12M posts, posted on Gab between August 10, 2016 and July 31, 2017, and 955K posts have at least one image, using the same methodology as in [72]. Out of these, 235K images are unique, producing 193K unique pHashes.

Ethics. Although we only collect publicly available data, our study has been approved by the designated ethics officer at UCL. Since 4chan content is typically posted with expectations of anonymity, we have encrypted data at rest, while making no attempt to deanonymize users.

3.2 Meme Annotation Site

Know Your Meme (KYM). We choose KYM as the source for meme annotation as it offers a comprehensive database of memes. KYM is a sort of encyclopedia of Internet memes: for each meme, it provides information such as its origin (i.e., the platform on which it was first observed), the year it started, as well as descriptions and examples. In addition, for each entry, KYM provides a set of keywords, called *tags*, that describe the entry. KYM provides a variety of higher-level categories that group meme entries; namely, cultures, subcultures, people, events, and sites. "Cultures" and "subcultures" entries refer to a wide variety of topics ranging from video games to various general categories. For example, the Rage Comics *subculture* [42] is a higher level category associated with memes

[1]See [64] for metadata associated with submissions and comments.



(a) Images (b) Origins

Figure 3: Basic statistics from the KYM dataset. We omit some categories (i.e., Sites, Events, and Subcultures) from Fig. 3(a) for readability purposes.

related to comics like Rage Guy [43] or LOL Guy [34], while the Alt-right *culture* [22] gathers entries from a loosely defined segment of the right-wing community. The rest of the categories refer to specific individuals (e.g., Donald Trump [29]), specific *events (e.g.,#CNNBlackmail [27]), and* sites (e.g., /pol/ [41]), respectively. It is also worth noting that KYM moderates all entries, hence entries that are wrong or incomplete are marked as so by the site.

As of May 2018, the site has 18.3K entries, specifically, 14K memes, 1.3K subcultures, 1.2K people, 1.3K events, and 427 websites [35]. We crawl KYM between October and December 2017, acquiring data for 15.6K entries; for each entry, we also download all the images related to it by crawling all the pages of the image gallery. In total, we collect 707K images corresponding to 597K unique pHashes. Note that we obtain 15.6K out of 18.3K entries, as we crawled the site several months before May 2018.

Getting to know KYM. We also perform a general characterization of KYM. First, we look at the distribution of entries across categories: as expected, the majority (57%) are memes, followed by subcultures (30%), cultures (3%), websites (2%), and people (2%). Next, we measure the number of images per entry: as shown in Fig. 3(a), this varies considerably (note log-scale on x-axis). KYM entries have as few as 1 and as many as 8K images, with an average of 45 and a median of 9 images. Larger values may be related to the meme's popularity, but also to the "diversity" of image variants it generates. Upon manual inspection, we find that the presence of a large number of images for the same meme happens either when images are visually very similar to each other (e.g., Smug Frog images *within* the two clusters in Fig. 1), or if there are actually remarkably different variants of the same meme (e.g., images in 'cluster 1' *vs.* images in 'cluster N' in the same figure). We also note that the distribution varies according to the category: e.g., higher-level concepts like cultures include more images than more specific entries like memes.

We then analyze the origin of each entry: see Fig. 3(b). Note that a large portion of the memes (28%) have an unknown origin, while YouTube, 4chan, and Twitter are the most popular platforms with, respectively, 21%, 12%, and 11%, followed by Tumblr and Reddit with 8% and 7%. This confirms our intuition that 4chan, Twitter, and Reddit, which are among our data sources, play an important role in the generation and dissemination of memes. As mentioned, we do not currently study video memes originating from YouTube, due to the inherent complexity of video-processing tasks as well as scalability issues. However, a large portion of YouTube memes

Platform	#Images	Noise	#Clusters	#Clusters with KYM tags (%)
/pol/	4,325,648	63%	38,851	9,265 (24%)
T_D	1,234,940	64%	21,917	2,902 (13%)
Gab	235,222	69%	3,083	447 (15%)

Table 2: Statistics obtained from clustering images from /pol/, The_Donald, and Gab.

actually end up being morphed into image-based memes (see, e.g., the Overly Attached Girlfriend meme [39]).

3.3 Running the pipeline on our datasets

For all four Web communities (Twitter, Reddit, /pol/, and Gab), we perform Step 1 of the pipeline (Fig. 2), using the ImageHash library [2]. We then perform Steps 2-3 (i.e., pairwise comparisons between all images and clustering), for all the images from /pol/, The_Donald subreddit, and Gab, as we treat them as fringe Web communities. Note that, we exclude mainstream communities like the rest of Reddit and Twitter as our main goal is to obtain clusters of memes from fringe Web communities and later characterize all communities by means of the clusters. Next, we go through Steps 4-5 using all the images obtained from meme annotation websites (specifically, Know Your Meme, see Sec. 3.2) and the medoid of each cluster from /pol/, The_Donald, and Gab. Finally, Steps 6-7 use all the pHashes obtained from Twitter, Reddit (all subreddits), /pol/, and Gab to find posts with images matching the annotated clusters. This is an integral part of our process as it allows to characterize and study mainstream Web communities not used for clustering (i.e., Twitter and Reddit).

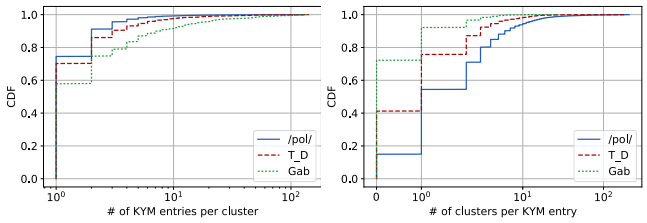
4 ANALYSIS

In this section, we present a cluster-based measurement of memes as well as an analysis of a few Web communities from the "perspective" of memes. We measure the prevalence of memes across the clusters obtained from fringe communities: /pol/, The_Donald subreddit (T_D), and Gab. We also use the distance metric introduced in Eq. 1 to perform a *cross-community* analysis. Then, we group clusters into broad, but related, categories to gain a macro-perspective understanding of larger communities, including mainstream ones like Reddit and Twitter.

4.1 Cluster-based Analysis

We start by analyzing the 12.6K annotated clusters consisting of 268K images from /pol/, The_Donald, and Gab (Step 5 in Fig. 2). We do so to understand the *diversity* of memes in each Web community, as well as the interplay between *variants* of memes. We then evaluate how clusters can be grouped into higher structures using hierarchical clustering and graph visualization techniques.

4.1.1 Clusters. **Statistics.** In Table 2, we report some basic statistics of the clusters obtained for each Web community. A relatively high percentage of images (63%–69%) are not clustered, i.e., are labeled as noise. While in DBSCAN "noise" is just an instance that does not fit in any cluster (more specifically, there are less than 5 images with perceptual distance ≤ 8 from that particular instance), we note that this likely happens as these images are not memes,



(a) (b)

Figure 4: CDF of KYM entries per cluster (a) and clusters per KYM entry (b).

but rather "one-off images." For example, on /pol/ there is a large number of pictures of random people taken from various social media platforms.

Overall, we have 2.1M images in 63.9K clusters: 38K clusters for /pol/, 21K for The_Donald, and 3K for Gab. 12.6K of these clusters are successfully annotated using the KYM data: 9.2K from /pol/ (142K images), 2.9K from The_Donald (121K images), and 447 from Gab (4.5K images). Examples of clusters are reported in Appendix D of the extended version of the paper [73]. As for the un-annotated clusters, manual inspection confirms that many include miscellaneous images unrelated to memes, e.g., similar screenshots of social networks posts (recall that we only filter out screenshots from the KYM image galleries), images captured from video games, etc.

KYM entries per cluster. Each cluster may receive multiple annotations, depending on the KYM entries that have at least one image matching that cluster's medoid. As shown in Fig. 4(a), the majority of the annotated clusters (74% for /pol/, 70% for The_Donald, and 58% for Gab) only have a single matching KYM entry. However, a few clusters have a large number of matching entries, e.g., the one matching the Conspiracy Keanu meme [28] is annotated by 126 KYM entries (primarily, other memes that add text in an image associated with that meme). This highlights that memes do overlap and that some are highly influenced by other ones.

Clusters per KYM entry. We also look at the number of clusters annotated by the *same* KYM entry. Fig. 4(b) plots the CDF of the number of clusters per entry. About 40% only annotate a single /pol/ cluster, while 34% and 20% of the entries annotate a single The_Donald and a single Gab cluster, respectively. We also find that a small number of entries are associated to a large number of clusters: for example, the Happy Merchant meme [32] annotates 124 different clusters on /pol/. This highlights the *diverse* nature of memes, i.e., memes are mixed and matched, not unlike the way that genetic traits are combined in biological reproduction.

Top KYM entries. Because the majority of clusters match only one or two KYM entries (Fig. 4(a)), we simplify things by giving all clusters a *representative annotation* based on the most prevalent annotation given to the medoid, and, in the case of ties the average distance between all matches (see Sec. 2.2). *Thus, in the rest of the paper, we report our findings based on the representative annotation for each cluster.*

In Table 3, we report the top 20 KYM entries with respect to the number of clusters they annotate. These cover 17%, 23%, and 27% of the clusters in /pol/, The_Donald, and Gab, respectively,

/pol/			T_D			Gab		
Entry	**Category**	**Clusters (%)**	**Entry**	**Category**	**Clusters (%)**	**Entry**	**Category**	**Clusters (%)**
Donald Trump	People	207 (2.2%)	Donald Trump	People	177 (6.1%)	Donald Trump	People	25 (5.6%)
Happy Merchant	Memes	124 (1.3%)	Smug Frog	Memes	78 (2.7%)	Happy Merchant	Memes	10 (2.2%)
Smug Frog	Memes	114 (1.2%)	Pepe the Frog	Memes	63 (2.1%)	Demotivational Posters	Memes	7 (1.5%)
Computer Reaction Faces	Memes	112 (1.2%)	Feels Bad Man/ Sad Frog	Memes	61 (2.1%)	Pepe the Frog	Memes	6 (1.3%)
Feels Bad Man/ Sad Frog	Memes	94 (1.0%)	Make America Great Again	Memes	50 (1.7%)	#Cnnblackmail	Events	6 (1.3%)
I Know that Feel Bro	Memes	90 (1.0%)	Bernie Sanders	People	31 (1.0%)	2016 US election	Events	6 (1.3%)
Tony Kornheiser's Why	Memes	89 (1.0%)	2016 US Election	Events	27 (0.9%)	Know Your Meme	Sites	6 (1.3%)
Bait/This is Bait	Memes	84 (0.9%)	Counter Signal Memes	Memes	24 (0.8%)	Tumblr	Sites	6 (1.3%)
#TrumpAnime/Rick Wilson	Events	76 (0.8%)	#Cnnblackmail	Events	24 (0.8%)	Feminism	Cultures	5 (1.1%)
Reaction Images	Memes	73 (0.8%)	Know Your Meme	Sites	20 (0.7%)	Barack Obama	People	5 (1.1%)
Make America Great Again	Memes	72 (0.8%)	Angry Pepe	Memes	18 (0.6%)	Smug Frog	Memes	5 (1.1%)
Counter Signal Memes	Memes	72 (0.8%)	Demotivational Posters	Memes	18 (0.6%)	rwby	Subcultures	5 (1.1%)
Pepe the Frog	Memes	65 (0.7%)	4chan	Sites	16 (0.5%)	Kim Jong Un	People	5 (1.1%)
Spongebob Squarepants	Subcultures	61 (0.7%)	Tumblr	Sites	15 (0.5%)	Murica	Memes	5 (1.1%)
Doom Paul its Happening	Memes	57 (0.6%)	Gamergate	Events	15 (0.5%)	UA Passenger Removal	Events	5 (1.1%)
Adolf Hitler	People	56 (0.6%)	Colbertposting	Memes	15 (0.5%)	Make America Great Again	Memes	4 (0.9%)
pol	Sites	53 (0.6%)	Donald Trump's Wall	Memes	15 (0.5%)	Bill Nye	People	4 (0.9%)
Dubs Guy/Check'em	Memes	53 (0.6%)	Vladimir Putin	People	15 (0.5%)	Trolling	Cultures	4 (0.9%)
Smug Anime Face	Memes	51 (0.6%)	Barack Obama	People	15 (0.5%)	4chan	Sites	4 (0.9%)
Warhammer 40000	Subcultures	51 (0.6%)	Hillary Clinton	People	15 (0.5%)	Furries	Cultures	3 (0.7%)
Total		1,638 (17.7%)			695 (23.9%)			121 (27.1%)

Table 3: Top 20 KYM entries appearing in the clusters of /pol/, The_Donald, and Gab. We report the number of clusters and their respective percentage (per community). Each item contains a hyperlink to the corresponding entry on the KYM website.

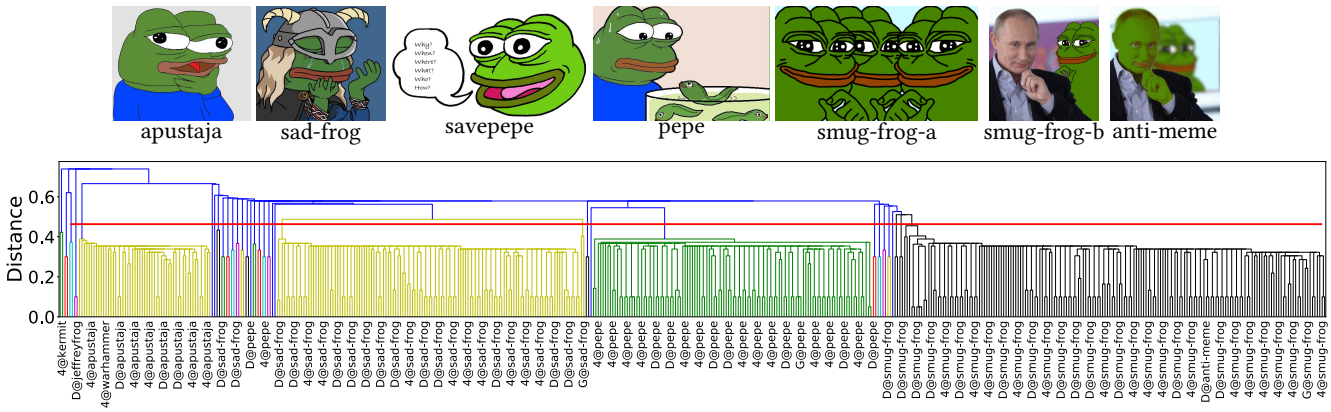


Figure 5: Inter-cluster distance between all clusters with frog memes. Clusters are labeled with the origin (*4* for 4chan, *D* for The_Donald, and *G* for Gab) and the meme name. To ease readability, we do not display all labels, abbreviate meme names, and only show an excerpt of all relationships.

hence covering a relatively good sample of our datasets. Donald Trump [29], Smug Frog [47], and Pepe the Frog [40] appear in the top 20 for all three communities, while the Happy Merchant [32] only in /pol/ and Gab. In particular, Donald Trump annotates the most clusters (207 in /pol/, 177 in The_Donald, and 25 in Gab). In fact, politics-related entries appear several times in the Table, e.g., Make America Great Again [36] as well as political personalities like Bernie Sanders, Obama, Putin, and Hillary Clinton.

When comparing the different communities, we observe the most prevalent categories are memes (6 to 14 entries in each community) and people (2-5). Moreover, in /pol/, the 2nd most popular entry, related to people, is Adolf Hilter, which supports previous reports of the community's sympathetic views toward Nazi ideology [20]. Overall, there are several memes with hateful or disturbing content (e.g., holocaust). This happens to a lesser extent in The_Donald and Gab: the most popular people after Donald Trump are contemporary politicians.

Finally, image posting behavior in fringe Web communities is greatly influenced by real-world events. For instance, in /pol/, we find the #TrumpAnime controversy event [50], where a political individual (Rick Wilson) offended the alt-right community, Donald Trump supporters, and anime fans (an oddly intersecting set of interests of /pol/ users). Similarly, on The_Donald and Gab, we find the #Cnnblackmail [27] event, referring to the (alleged) blackmail of the Reddit user that created the infamous video of Donald Trump wrestling the CNN.

4.1.2 Memes' Branching Nature. Next, we study how memes *evolve* by looking at *variants* across different clusters. Intuitively, clusters that look alike and/or are part of the same meme are grouped together under the same branch of an evolutionary tree. We use the

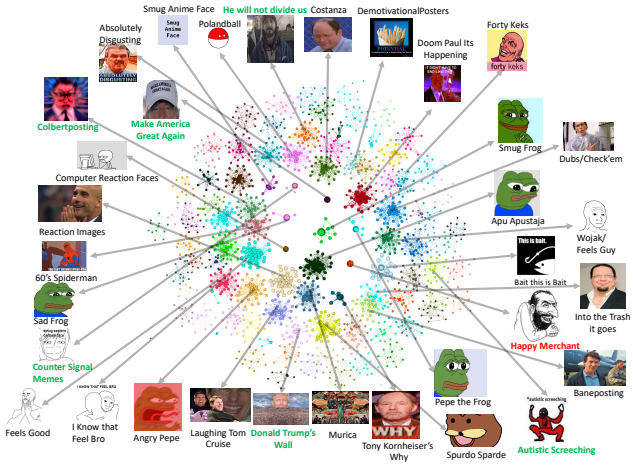


Figure 6: Visualization of the obtained clusters from /pol/, The_Donald, and Gab. Note that memes with red labels are annotated as racist, while memes with green labels are annotated as politics (see Section 4.2.1 for the selection criteria).

custom distance metric introduced in Sec. 2.3, aiming to infer the phylogenetic relationship between variants of memes. Since there are 12.6K annotated clusters, we only report on a subset of variants. In particular, we focus on "frog" memes (e.g., Pepe the Frog [40]); as discussed later in Sec. 4.2, this is one of the most popular memes in our datasets.

The dendrogram in Fig. 5 shows the hierarchical relationship between groups of clusters of memes related to frogs. Overall, there are 525 clusters of frogs, belonging to 23 different memes. These clusters can be grouped into four large categories, dominated by different memes that express different ideas or messages: e.g., Apu Apustaja depicts a simple-minded non-native speaker using broken English, while the Feels Bad Man/Sad Frog (ironically) expresses dismay at a given situation. The dendrogram also shows a variant of Smug Frog (*smug-frog-b*) related to a variant of the Russian Anti Meme Law [46] (*anti-meme*) as well as relationships between clusters from Pepe the Frog and Isis meme [33], and between Smug Frog and Brexit-related clusters [51], as shown in Appendix E in [73].

The distance metric quantifies the similarity of any two variants of *different* memes; however, recall that two clusters can be close to each other even when the medoids are perceptually different (see Sec. 2.3), as in the case of Smug Frog variants in the *smug-frog-a* and *smug-frog-b* clusters (top of Fig. 5). Although, due to space constraints, this analysis is limited to a single "family" of memes, our distance metric can actually provide useful insights regarding the phylogenetic relationships of any clusters. In fact, more extensive analysis of these relationships (through our pipeline) can facilitate the understanding of the diffusion of ideas and information across the Web, and provide a rigorous technique for large-scale analysis of Internet culture.

4.1.3 Meme Visualization. We also use the custom distance metric (see Eq. 1) to visualize the clusters with annotations. We build a graph $G = (V, E)$, where V are the medoids of annotated clusters and E the connections between medoids with distance under a threshold κ. Fig. 6 shows a snapshot of the graph for $\kappa = 0.45$, chosen based on the frogs analysis above (see red horizontal line in Fig. 5). In particular, we select this threshold as the majority of the clusters from the same meme (note coloration in Fig. 5) are hierarchically connected with a higher-level cluster at a distance close to 0.45. To ease readability, we filter out nodes and edges that have a sum of in- and out-degree less than 10, which leaves 40% of the nodes and 92% of the edges. Nodes are colored according to their KYM annotation. NB: the graph is laid out using the OpenOrd algorithm [58] and the distance between the components in it does not exactly match the actual distance metric. We observe a large set of disconnected components, with each component containing nodes of primarily one color. This indicates that our distance metric is indeed capturing the peculiarities of different memes. Finally, note that an interactive version of the full graph is publicly available from [1].

4.2 Web Community-based Analysis

We now present a macro-perspective analysis of the Web communities through the lens of memes. We assess the presence of different memes in each community, how popular they are, and how they evolve. To this end, we examine the *posts* from all four communities (Twitter, Reddit, /pol/, and Gab) that contain *images* matching *memes* from fringe Web communities (/pol/, The_Donald, and Gab).

4.2.1 Meme Popularity. We start by analyzing clusters grouped by KYM 'meme' entries, looking at the number of posts for each meme in /pol/, Reddit, Gab, and Twitter. (We also include the analysis for 'people' entries in the extended version [73].)

In Table 4, we report the top 20 memes for each Web community sorted by the number of posts. We observe that Pepe the Frog [40] and its variants are among the most popular memes for every platform. While this might be an artifact of using fringe communities as a "seed" for the clustering, recall that the goal of this work is in fact to gain an understanding of how fringe communities disseminate memes and influence mainstream ones. Thus, we leave to future work a broader analysis of the wider meme ecosystem.

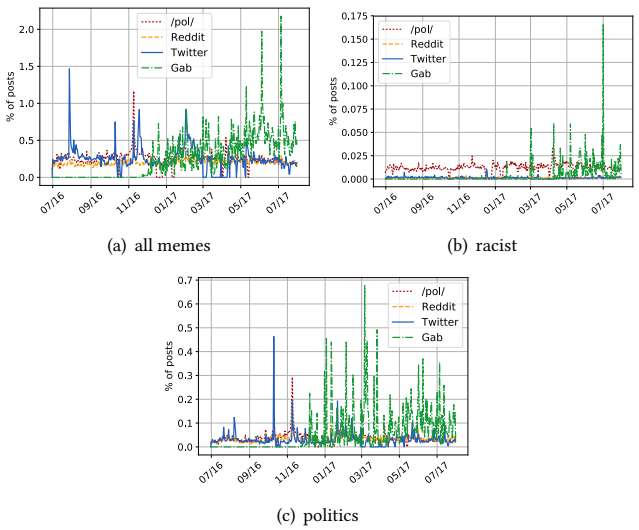
Sad Frog [31] is the most popular meme on /pol/ (4.9%), the 3rd on Reddit (1.3%), the 10th on Gab (0.8%), and the 12th on Twitter (0.5%). We also find variations like Smug Frog [47], Apu Apustaja [24], Pepe the Frog [40], and Angry Pepe [23]. Considering that Pepe is treated as a hate symbol by the Anti-Defamation League [53] and that is often used in hateful or racist, this likely indicates that polarized communities like /pol/ and Gab do use memes to incite hateful conversation. This is also evident from the popularity of the anti-semitic Happy Merchant meme [32], which depicts a "greedy" and "manipulative" stereotypical caricature of a Jew (3.8% on /pol/ and 1.1% on Gab).

By contrast, mainstream communities like Reddit and Twitter primarily share harmless/neutral memes, which are rarely used in hateful contexts. Specifically, on Reddit the top memes are Manning Face [37] (2.2%) and That's the Joke [48] (1.3%), while on Twitter the top ones are Roll Safe [45] (5.9%) and Evil Kermit [30] (5.4%).

Once again, we find that users (in all communities) post memes to share politics-related information, possibly aiming to enhance or penalize the public image of politicians (see Appendix E of the paper's extended version [73] for an example of such memes). For instance, we find Make America Great Again [36], a meme dedicated

/pol/		Reddit		Gab		Twitter	
Entry	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts(%)**
Feels Bad Man/Sad Frog	64,367 (4.9%)	Manning Face	12,540 (2.2%)	Jesusland **(P)**	454 (1.6%)	Roll Safe	55,010 (5.9%)
Smug Frog	63,290 (4.8%)	That's the Joke	7,626 (1.3%)	Demotivational Posters	414 (1.5%)	Evil Kermit	50,642 (5.4%)
Happy Merchant **(R)**	49,608 (3.8%)	Feels Bad Man/ Sad Frog	7,240 (1.3%)	Smug Frog	392 (1.4%)	Arthur's Fist	37,591 (4.0%)
Apu Apustaja	29,756 (2.2%)	Confession Bear	7,147 (1.3%)	Based Stickman **(P)**	391 (1.4%)	Nut Button	13,598 (1,5%)
Pepe the Frog	25,197 (1.9%)	This is Fine	5,032 (0.9%)	Pepe the Frog	378 (1.3%)	Spongebob Mock	11,136 (1,2%)
Make America Great Again **(P)**	21,229 (1.6%)	Smug Frog	4,642 (0.8%)	Happy Merchant **(R)**	297 (1.1%)	Reaction Images	9,387 (1.0%)
Angry Pepe	20,485 (1.5%)	Roll Safe	4,523 (0.8%)	Murica	274 (1.0%)	Conceited Reaction	9,106 (1.0%)
Bait this is Bait	16,686 (1.2%)	Rage Guy	4,491 (0.8%)	And Its Gone	235 (0.9%)	Expanding Brain	8,701 (0.9%)
I Know that Feel Bro	14,490 (1.1%)	Make America Great Again **(P)**	4,440 (0.8%)	Make America Great Again **(P)**	207 (0.8%)	Demotivational Posters	7,781 (0.8%)
Cult of Kek	14,428 (1.1%)	Fake CCG Cards	4,438 (0.8%)	Feels Bad Man/ Sad Frog	206 (0.8%)	Cash Me Ousside/Howbow Dah	5,972 (0.6%)
Laughing Tom Cruise	14,312 (1.1%)	Confused Nick Young	4,024 (0.7%)	Trump's First Order of Business **(P)**	192 (0.7%)	Salt Bae	5,375 (0.6%)
Awoo	13,767 (1.0%)	Daily Struggle	4,015 (0.7%)	Kekistan	186 (0.6%)	Feels Bad Man/ Sad Frog	4,991 (0.5%)
Tony Kornheiser's Why	13,577 (1.0%)	Expanding Brain	3,757 (0.7%)	Picardia **(P)**	183 (0.6%)	Math Lady/Confused Lady	4,722 (0.5%)
Picardia **(P)**	13,540 (1.0%)	Demotivational Posters	3,419 (0.6%)	Things with Faces (Pareidolia)	156 (0.5%)	Computer Reaction Faces	4,720 (0.5%)
Big Grin / Never Ever	12,893 (1.0%)	Actual Advice Mallard	3,293 (0.6%)	Serbia Strong/Remove Kebab	149 (0.5%)	Clinton Trump Duet **(P)**	3,901 (0.4%)
Reaction Images	12,608 (0.9%)	Reaction Images	2,959 (0.5%)	Riot Hipster	148 (0.5%)	Kendrick Lamar Damn Album Cover	3,656 (0.4%)
Computer Reaction Faces	12,247 (0.9%)	Handsome Face	2,675 (0.5%)	Colorized History	144 (0.5%)	What in tarnation	3,363 (0.3%)
Wojak / Feels Guy	11,682 (0.9%)	Absolutely Disgusting	2,674 (0.5%)	Most Interesting Man in World	140 (0.5%)	Harambe the Gorilla	3,164 (0.3%)
Absolutely Disgusting	11,436 (0.8%)	Pepe the Frog	2,672 (0.5%)	Chuck Norris Facts	131 (0.4%)	I Know that Feel Bro	3,137 (0.3%)
Spurdo Sparde	9,581 (0.7%)	Pretending to be Retarded	2,462 (0.4%)	Roll Safe	131 (0.4%)	This is Fine	3,094 (0.3%)
Total	445,179 (33.4%)		94,069 (16.7%)		4,808 (17.0%)		249,047 (26.4%)

Table 4: Top 20 KYM entries for memes that we find our datasets. We report the number of posts for each meme as well as the percentage over all the posts (per community) that contain images that match one of the annotated clusters. The (R) and (P) markers indicate whether a meme is annotated as racist or politics-related, respectively (see Section 4.2.1 for the selection criteria).



(a) all memes (b) racist

(c) politics

Figure 7: Percentage of posts per day in our dataset for all, racist, and politics-related memes.

to Donald Trump's US presidential campaign, among the top memes in /pol/ (1.6%), in Reddit (0.8%), and Gab (0.8%). Similarly, in Twitter, we find the Clinton Trump Duet meme [26] (0.4%), a meme inspired by the 2nd US presidential debate.

We further group memes into two high-level groups, racist and politics-related. We use the *tags* that are available in our KYM dataset, i.e., we assign a meme to the politics-related group if it has the "politics," "2016 us presidential election," "presidential election," "trump," or "clinton" tags, and to the racism-related one if the tags include "racism," "racist," or "antisemitism," obtaining 117 racist memes(4.4% of all memes that appear on our dataset) and 556 politics-related memes (21.2% of all memes that appear on our dataset). In the rest of this section, we use these groups to further study the memes, and later in Sec. 5 to estimate influence.

4.2.2 Temporal Analysis. Next, we study the temporal aspects of posts that contain memes from /pol/, Reddit, Twitter, and Gab. In Fig. 7, we plot the percentage of posts per day that include memes. For all memes (Fig. 7(a)), we observe that /pol/ and Reddit follow a steady posting behavior, with a peak in activity around the 2016 US elections. We also find that memes are increasingly more used on Gab (see, e.g., 2016 vs 2017).

As shown in Fig. 7(b), both /pol/ and Gab include a substantially higher number of posts with racist memes, used over time with a difference in behavior: while /pol/ users share them in a very steady and constant way, Gab exhibits a bursty behavior. A possible explanation is that the former is inherently more racist, with the latter primarily reacting to particular world events. As for political memes (Fig. 7(c)), we find a lot of activity overall on Twitter, Reddit, and /pol/, but with different spikes in time. On Reddit and /pol/, the peaks coincide with the 2016 US elections. On Twitter, we note a peak that coincides with the 2nd US Presidential Debate on October 2016. For Gab, there is again an increase in posts with political memes after January 2017.

4.3 Take-Aways

In summary, the main take-aways of our analysis include:

(1) Fringe Web communities use many variants of memes related to politics and world events, possibly aiming to share weaponized information about them (see Appendix E in [73] for some examples of weaponized memes). For instance, Donald Trump is the KYM entry with the largest number of clusters in /pol/ (2.2%), The_Donald (6.1%), and Gab (2.2%).
(2) /pol/ and Gab share hateful and racist memes at a higher rate than mainstream communities, as we find a considerable number of anti-semitic and pro-Nazi clusters (e.g., The Happy Merchant meme [32] appears in 1.3% of all /pol/ annotated clusters and 2.2% of Gab's, while Adolf Hitler in 0.6% of /pol/'s). This trend is steady over time for /pol/ but ramping up for Gab.

(3) Seemingly "neutral" memes, like Pepe the Frog (or one of its variants), are used in conjunction with other memes to incite hate or influence public opinion on world events, e.g., with images related to terrorist organizations like ISIS or world events such as Brexit.
(4) Our custom distance metric successfully allows us to study the interplay and the overlap of memes, as showcased by the visualizations of the clusters and the dendrogram (see Figs. 5 and 6).

5 INFLUENCE ESTIMATION

So far we have studied the dissemination of memes by looking at Web communities in isolation. However, in reality, these influence each other: e.g., memes posted on one community are often re-posted to another. Aiming to capture the relationship between them, we use a statistical model known as Hawkes Processes [56, 57], which describes how events occur over time on a collection of processes. This maps well to the posting of memes on different platforms: each community can be seen as a process, and an event occurs each time a meme image is posted on one of the communities. Events on one process can cause impulses that can increase the likelihood of subsequent events, including other processes, e.g., a person might see a meme on one community and re-post it, or share it to a different one. This approach allow us to assess the causality of events, hence it is a far better approach when compared to simple approaches like looking at the timeline of specific memes or phashes.

5.1 Hawkes Processes

To model the spread of memes on Web communities, we use a similar approach as in our previous work [74], which looked at the spread of mainstream and alternative news URLs. Next, we provide a brief description, and present an improved method for estimating influence.

We use five processes, one for each of our seed Web communities (/pol/, Gab, and The_Donald), as well as Twitter and Reddit, fitting a separate model for each meme cluster. Fitting the model to the data yields a number of values: background rates for each process, weights from each process to each other, and the shape of the impulses an event on one process causes on the rates of the others. The background rate is the expected rate at which events will occur on a process *without* influence from the communities modeled or previous events; this captures memes posted for the first time, or those seen on a community we do not model and then reposted on a community we do. The weights from community-to-community indicate the effect an event on one has on the others; for example, a weight from Twitter to Reddit of 1.2 means that each event on Twitter will cause an expected 1.2 additional events on Reddit. The shape of the impulse from Twitter to Reddit determines how the probability of these events occurring is distributed over time; typically the probability of another event occurring is highest soon after the original event and decreases over time.

Fig. 8 illustrates a Hawkes model with three processes. The first event occurs on process B, which causes an increase in the rate of events on all three processes. The second event then occurs on process C, again increasing the rate of events on the processes.

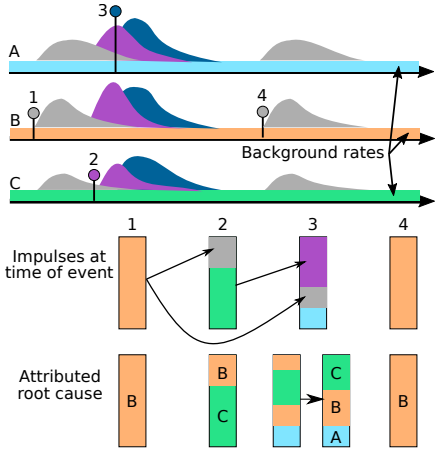


Figure 8: Representation of a Hawkes model with three processes. Events cause impulses that increase the rate of subsequent events in the same or other processes. By looking at the impulses present when events occur, the probability of a process being the root cause of an event can be determined. Note that on the second part of the Figure, colors represent events while arrows represent impulses between the events.

The third event occurs soon after, on process A. The fourth event occurs later, again caused by the background arrival rate on process B, after the increases in arrival rate from the other events have disappeared.

To understand the influence different communities have on the spread of memes, we want to be able to attribute the cause of a meme being posted back to a specific community. For example, if a meme is posted on /pol/ and then someone sees it there and posts it on Twitter where it is shared several times, we would like to be able to say that /pol/ was the *root cause* of those events. Obviously, we do not actually know where someone saw something and decided to share it, but we can, using the Hawkes models, determine the *probability* of each community being the root cause of an event.

Looking again at Fig. 8, we see that events 1 and 4 are caused directly by the background rate of process B. This is because, in the case of event 1, there are no previous events on other processes, and in the case of event 4, the impulses from previous events have already stopped. Events 2 and 3, however, occur when there are multiple possible causes: the background rate for the community and the impulses from previous events. In these cases, we assign the probability of being the root cause in proportion to the magnitudes of the impulses (including the background rate) present at the time of the event. For event 2, the impulse from event 1 is smaller than the background rate of community C, so the background rate has a higher probability of being the cause of event 2 than event 1. Thus, most of the cause for event 2 is attributed to community C, with a lesser amount to B (through event 1). Event 3 is more complicated: impulses from both previous events are present, thus the probability of being the cause is split three ways, between the background rate and the two previous events. The impulse from event 2 is the largest, with the background rate and event 1 impulse smaller. Because event 2 is attributed both to communities B and C,

/pol/	Twitter	Reddit	T_D	Gab
1,574,045	865,885	581,803	81,924	44,918

Table 5: Events per community from the 12.6K clusters.

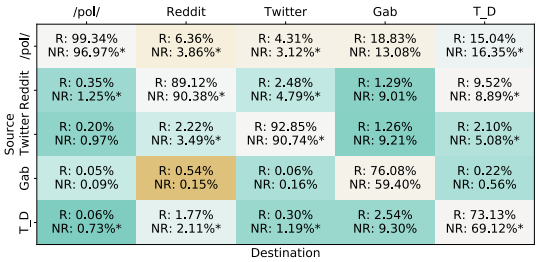


Figure 9: Percent of the destination community's racist (R) and non-racist (NR) meme postings caused by the source community. Colors indicate the percent difference between racist and non-racist.

event 3 is partly attributed to community B through both event 1 and event 2.

In the rest of our analysis, we use this new measure. This is a substantial improvement over the influence estimation in [74], which used the weights from source to destination community, multiplied by the number of events on the source to estimate influence. However, this only looks at influence across a single "hop" and would not allow us to understand the source community's influence as memes spread onwards from the destination community. The new method allows us to gain an understanding of where memes that appear on a community originally come from, and how they are likely to spread from community to community from the original source.

5.2 Influence

We fit Hawkes models using Gibbs sampling as described in [57] for the 12.6K annotated clusters; in Table 5, we report the total number of meme images posted to each community in these clusters. We note that /pol/ has the greatest number of memes posted, followed by Twitter and Reddit. Recall, however, that because our approach is seeded with memes observed on /pol/, The_Donald, and Gab, it is possible that there are memes on Twitter and Reddit that are not included in the clusters. In addition, the raw number of images (not necessarily memes) that appear on the different communities varies greatly (see Table 1). This yields an additional interesting question: how *efficient* are different communities at disseminating memes?

First, we report the source of events in terms of the percent of events on the destination community. This describes the results in terms of the data as we have collected it, e.g., it tells us the percentage of memes posted on Twitter that were caused by /pol/. The second way we report influence is by normalizing the values by the total number of events in the source community, which lets us see how much influence each community has, relative to the number of memes they post—in other words, their efficiency.

Additional results are available in extended version of the paper [73]; here we focus on the differences in the ways communities disseminate different *types* of meme, in particular, *racist* (non-racist) and *political* (non-political) memes.

Using the clusters identified as either racist or non-racist (see the end of Sec. 4.2.1), we compare how the communities influence the

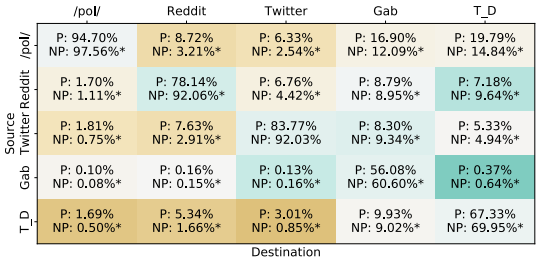


Figure 10: Percent of the destination community's political (P) and non-political (NP) meme postings caused by the source community. Colors indicate the percent difference between political and non-political.

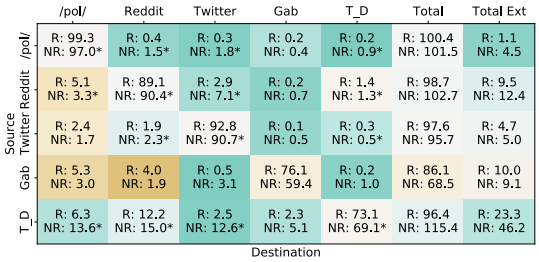


Figure 11: Influence from source to destination community of racist and non-racist meme postings, normalized by the number of events in the *source* community.

spread of these two types of content. Fig. 9 shows the percentage of both the destination community's racist and non-racist meme posts caused by the source community. We perform two-sample Kolmogorov-Smirnov tests to compare the distributions of influence from the racist and non-racist clusters; cells with statistically significant differences between influence of racist/non-racist memes (with $p<0.01$) are reported with a * in the figure. /pol/ has the most *total* influence for both racist and non-racist memes, with the notable exception of Twitter, where Reddit has the most the influence. Interestingly, while the percentage of racist meme posts caused by /pol/ is greater than non-racist for Reddit, Twitter, and Gab, this is *not* the case for The_Donald. The only other cases where influence is greater for racist memes are Reddit to The_Donald and Gab to Reddit.

When looking at political vs non political memes (Fig. 10), we see a somewhat different story. Here, /pol/ influences The_Donald more in terms of political memes. Further, we see differences in the *percent* increase and decrease of influence between the two figures (as indicated by the cell colors). For example, Twitter has a relatively larger difference in its influence on /pol/ and Reddit for political and non-political memes than for racist and non-racist memes, but a smaller difference in its influence on Gab and The_Donald. This exposes how different communities have varying levels of influence depending on the type of memes they post.

While examining the raw influence provides insights into the meme ecosystem, it obscures notable differences in the meme posting behavior of the different communities. To explore this, we look at the *normalized* influence in Fig. 11 (racist/non-racist memes) and Fig. 12 (political/non-political memes). As mentioned previously, normalization reveals how *efficient* the communities are in disseminating memes to other communities by revealing the *per meme*

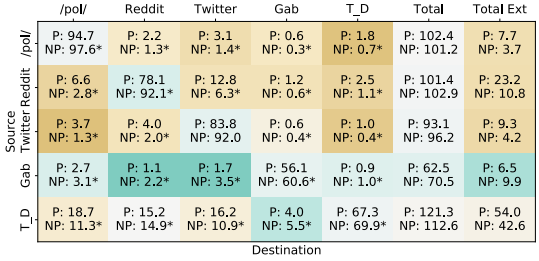


Figure 12: Influence from source to destination community of political and non-political meme postings, normalized by the number of events in the *source* community.

influence of meme posts. First, we note that the percent change in influence for the dissemination of racist/non-racist memes is quite a bit larger than that for political/non-political memes (again, indicated by the coloring of the cells). More interestingly, both figures show that, contrary to the *total* influence, /pol/ is the *least* influential when taking into account the number of memes posted. While this might seem surprising, it actually yields a subtle, yet crucial aspect of /pol/'s role in the meme ecosystem: /pol/ (and 4chan in general) acts as an evolutionary microcosm for memes. The constant production of new content [20] results in a "survival of the fittest" [15] scenario. A staggering number of memes are posted on /pol/, but only the *best* actually make it out to other communities. To the best of our knowledge, this is the first result quantifying this analogy to evolutionary pressure.

Take-Aways. There are several take-aways from our measurement of influence. We show that /pol/ is, generally speaking, the most influential disseminator of memes in terms of raw influence, In particular, it is more influential in spreading *racist* memes than non-racist one, and this difference is deeper than in any other community. There is one notable exception: /pol/ is more influential in terms of *non-racist* memes on The_Donald. Relatedly, /pol/ has generally more influence in terms of spreading political memes than other communities. When looking at the normalized influence, however, we surface a more interesting result: /pol/ is the *least* efficient in terms of influence while The_Donald is the *most* efficient. This provides new insight into the meme ecosystem: there are clearly evolutionary effects. Many meme postings do not result in further dissemination, and one of the key components to ensuring they are disseminated is ensuring that new "offspring" are continuously produced. /pol/'s "famed" meme magic, i.e., the propensity to produce and heavily push memes, is thus the most likely explanation for /pol/'s influence on the Web in general.

6 RELATED WORK

Detection and Propagation of Memes. Leskovec et al. [55] perform large-scale tracking of text-based memes, focusing on news outlets and blogs. Ferrara et al. [16] detect text memes using an unsupervised framework based on clustering techniques. Ratkiewicz et al. [63] introduce Truthy, a framework supporting the analysis of the diffusion of text-based, politics-related memes on Twitter. Babaei et al. [7] study Twitter users' preferences, with respect to information sources, including how they acquire memes. Dubey et al. [13] extract a rich semantic embedding corresponding to the template used for the creation of meme images. By contrast, we focus on the detection and propagation of image-based memes without limiting our scope to image macros. We also detect and study the propagation of memes across multiple Web communities, using publicly available data from memes annotation sites (i.e., KYM).

Popularity of Memes. Weng et al. [68, 69] study the popularity of memes spreading as hashtags on Twitter. They model virality using an agent-based approach, taking into account that users have a limited capacity in receiving/viewing memes on Twitter. They study the features that make memes popular, finding that those based on network community structures are strong indicators of popularity. Tsur and Rappoport [67] predict popularity of text-based memes on Twitter using linguistic characteristics as well as cognitive and domain features. Ienco et al. [21] study memes propagating via text, images, audio, and video on the Yahoo! Meme platform (a platform discontinued in 2012), aiming to predict virality and select memes to be shown to users after login.

Whereas, we also study the popularity of memes, however, unlike previous work, we rely on a multi-platform approach, encompassing data from /pol/, Reddit, Twitter, and Gab, and show that the popularity of memes depends on the Web community and its ideology. For instance, /pol/ is well-known for its anti-semitic ideology and in fact the "Happy Merchant" meme [32] is the 3rd most popular meme on /pol/.

Evolution of Memes. Adamic et al. [6] study the evolution of text-based memes on Facebook, showing that it can be modeled by the Yule process. They find that memes significantly evolve and new variants appear as they propagate, and that specific communities within the network diffuse specific variants of a meme. Bauckhage [8] study the temporal dynamics of 150 memes using data from Google Insights as well as social bookmarking services like Digg, showing that different communities exhibit different interests/behaviors for different memes, and that epidemiology models can be used to predict the evolution and popularity of memes. By contrast, we study the temporal aspect of memes using Hawkes processes. This statistical framework allows us to assess the causality of the posting of memes on various Web communities, thus modeling their evolution and their influence across multiple communities.

Case Studies. Heath et al. [19] present a case study of how people perceive memes with a focus on urban legends, finding that they are more willing to share memes that evoke stronger disgust. Xie et al. [70] focus on YouTube memes, performing a large-scale keyword-based search for videos related to the Iranian election in 2009, extracting frequently used images and video segments. They show that most of the videos are not original, thus, meme-related techniques can be exploited to deduplicate content and capture the content diffusion on the Web. Finally, Dewan et al. [12] study the sentiment and content of images that are disseminated during crisis events like the 2015 Paris terror attacks. They analyze 57K images related to the attacks, finding instances of misinformation and conspiracy theories.

We also present a case study focusing on image memes of Pepe the Frog (see the discussion about Fig. 5). This showcases both the overlap and the diversity of certain memes, as well as how memes can be influenced by real-world events, with new variants being generated. For instance, after the UK Brexit referendum in 2016,

memes with Pepe the Frog started to be used in the Brexit context (see Appendix E in [73]).

Fringe Communities. Previous work has also shed light on fringe Web communities like 4chan, Gab, and sub-communities within Reddit. Bernstein et al. [9] study the ephemerality and anonymity features of the 4chan community using data from the Random board (/b/). Hine et al. [20] focus on /pol/, analyzing 8M posts and detecting a high volume of hate speech as well as the phenomenon of "raids," i.e., coordinated attacks aimed at disrupting other services. Zannettou et al. [72] analyze 22M posts from 336K users on Gab, finding that hate speech occurs twice as much as in Twitter, but twice less than /pol/. Snyder et al. [66] measure doxing on 4chan and 8chan, while Chandrasekharan et al. [10] introduce a computational approach to detect abusive content also looking at 4chan and Reddit.

Finally, Hawkes processes have also been used to quantify influence of fringe Web communities like /pol/ and The_Donald to mainstream ones like Twitter in the context of misinformation [74]. We follow a similar approach here, but use an improved method of determining the influence of the different communities.

7 DISCUSSION & CONCLUSION

In this paper, we presented a large-scale measurement study of the meme ecosystem. We introduced a novel image processing pipeline and ran it over 160M images collected from four Web communities (4chan's /pol/, Reddit, Twitter, and Gab). We clustered images from fringe communities (/pol/, Gab, and Reddit's The_Donald) based on perceptual hashing and a custom distance metric, annotated the clusters using data gathered from Know Your Meme, and analyzed them along a variety of axes. We then associated images from all the communities to the clusters to characterize them through the lens of memes and the influence they have on each other.

Our analysis highlights that the meme ecosystem is quite complex, with intricate relationships between different memes and their variants. We found important differences between the memes posted on different communities (e.g., Reddit and Twitter tend to post "fun" memes, while Gab and /pol/ racist or political ones). When measuring the influence of each community toward disseminating memes to other Web communities, we found that /pol/ has the largest overall influence for racist and political memes, however, /pol/ was the least *efficient*, i.e., in terms of influence w.r.t. the total number of memes posted, while The_Donald is very successful in pushing memes to both fringe and mainstream Web communities.

Our work constitutes the first attempt to provide a multi-platform measurement of the meme ecosystem, with a focus on fringe and potentially dangerous communities. Considering the increasing relevance of digital information on world events, our study provides a building block for future cultural anthropology work, as well as for building systems to protect against the dissemination of harmful ideologies. Moreover, our pipeline can already be used by social network providers to assist the identification of hateful content; for instance, Facebook is taking steps to ban Pepe the Frog used in the context of hate [54], and our methodology can help them automatically identify hateful variants. Finally, our pipeline can be used for tracking the propagation of images from any context or other language spheres, provided an appropriate annotation dataset.

Additional Results. Due to space constraints, we focused on the most salient results, leaving several to the extended version [73]. Some relevant "highlights" include: 1) Reddit posts including political memes receive higher scores compared to non-political ones, and those with racist memes lower scores than non-racist ones; 2) In general, Reddit users are more interested in politics-related memes than other types, while The_Donald is the most active subreddit overall in terms of all memes as well as racist and political ones; 3) A surprising number of racist memes appear in seemingly "neutral" communities, e.g., the AdviceAnimals subreddit; 4) Donald Trump is, by far, the most prevalent political figure in image posts across all four communities. Moreover, the extended version has appendices that provide additional details of our screenshot classifier as well as interesting instances of memes and meme clusters.

Performance. We also measured the time that it takes to associate images posted on Web communities to memes. All other steps in our system are one-time batch tasks, only executed if the annotations dataset is updated. To ease presentation, we only report the time to compare all the 74M images from Twitter (the largest dataset) against the medoids of all 12K annotated clusters: it took about 12 days on our infrastructure, equipped with two NVIDIA Titan Xp GPUs. This corresponds to 14ms per image, or 73 images per second. Note that, if new GPUs are added to our infrastructure, the workload would be divided equally across all GPUs.

Future work. In future work, we plan to include memes in video format, thus extending to other communities (e.g., YouTube). We also plan to study the *content* of the posts that contain memes, incorporating OCR techniques to capture associated text-based features that memes usually contain, and improving on KYM annotations via crowdsourced labeling.

Acknowledgments. We thank the anonymous reviewers and our shepherd Christo Wilson for their insightful feedback. This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025). We also gratefully acknowledge the support of the NVIDIA Corporation, for the donation of the two Titan Xp GPUs used for our experiments.

REFERENCES

[1] Graph visualization of the clusters. https://memespaper.github.io/.
[2] ImageHash Python Library. https://github.com/JohannesBuchner/imagehash.
[3] Paper Repository. https://github.com/memespaper/memes_pipeline.
[4] 4plebs. 4chan archive. http://4plebs.org/.
[5] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*, 2016.
[6] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng. Information Evolution in Social Networks. In *WSDM*, 2016.
[7] M. Babaei, P. Grabowicz, I. Valera, K. P. Gummadi, and M. Gomez-Rodriguez. On the Efficiency of the Information Networks in Social Media. In *WSDM*, 2016.
[8] C. Bauckhage. Insights into Internet Memes. In *ICWSM*, 2011.
[9] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *ICWSM*, 2011.
[10] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *CHI*, 2017.
[11] R. Dawkins. *The selfish gene.* Oxford university press, 1976.
[12] P. Dewan, A. Suri, V. Bharadhwaj, A. Mithal, and P. Kumaraguru. Towards Understanding Crisis Events On Online Social Networks Through Pictures. In *ASONAM*, 2017.

[13] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan. MemeSequencer: Sparse Matching for Embedding Image Macros. In *WWW*, 2018.
[14] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996.
[15] C. Faife. How 4Chan's Structure Creates a 'Survival of the Fittest' for Memes. https://motherboard.vice.com/en_us/article/ywzm8m/how-4chans-structure-creates-a-survival-of-the-fittest-for-memes, 2017.
[16] E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini. Clustering Memes in Social Media. In *ASONAM*, 2013.
[17] G. Graham. *Genes: A Philosophical Inquiry*. Routledge, 2005.
[18] D. Haddow. Meme warfare: how the power of mass replication has poisoned the US election. https://www.theguardian.com/us-news/2016/nov/04/political-memes-2016-election-hillary-clinton-donald-trump, 2016.
[19] C. Heath, C. Bell, and E. Sternberg. Emotional selection in memes: the case of urban legends. *Journal of Personality and Social Psychology*, 2001.
[20] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*, 2017.
[21] D. Ienco, F. Bonchi, and C. Castillo. The Meme Ranking Problem: Maximizing Microblogging Virality. In *ICDMW*, 2010.
[22] Know Your Meme. Alt-Right Culture. http://knowyourmeme.com/memes/cultures/alt-right, 2018.
[23] Know Your Meme. Angry Pepe Meme. http://knowyourmeme.com/memes/angry-pepe, 2018.
[24] Know Your Meme. Apu Apustaja Meme. http://knowyourmeme.com/memes/apu-apustaja, 2018.
[25] Know Your Meme. Bad Luck Brian Meme. http://knowyourmeme.com/memes/bad-luck-brian, 2018.
[26] Know Your Meme. Clinton/Trump Duet Meme. http://knowyourmeme.com/memes/make-america-great-again, 2018.
[27] Know Your Meme. #CNNBlackmail. http://knowyourmeme.com/memes/events/cnnblackmail, 2018.
[28] Know Your Meme. Conspiracy Keanu Meme. http://knowyourmeme.com/memes/conspiracy-keanu, 2018.
[29] Know Your Meme. Donald Trump Meme. http://knowyourmeme.com/memes/people/donald-trump, 2018.
[30] Know Your Meme. Evil Kermit Meme. http://knowyourmeme.com/memes/evil-kermit, 2018.
[31] Know Your Meme. Feels Bad Man / Sad Frog Meme. http://knowyourmeme.com/memes/feels-bad-man-sad-frog, 2018.
[32] Know Your Meme. Happy Merchant Meme. http://knowyourmeme.com/memes/happy-merchant, 2018.
[33] Know Your Meme. ISIS / Daesh Meme. http://knowyourmeme.com/memes/people/isis-daesh, 2018.
[34] Know Your Meme. LOL Guy Meme. http://knowyourmeme.com/memes/lol-guy, 2018.
[35] Know Your Meme. Main page of KYM entries. http://knowyourmeme.com/memes/, 2018.
[36] Know Your Meme. Make America Great Again Meme. http://knowyourmeme.com/memes/make-america-great-again, 2018.
[37] Know Your Meme. Manning Face Meme. http://knowyourmeme.com/memes/manningface, 2018.
[38] Know Your Meme. Maxvidya Meme. https://knowyourmeme.com/memes/maxvidya, 2018.
[39] Know Your Meme. Overly Attached Girlfriend Meme. http://knowyourmeme.com/memes/overly-attached-girlfriend, 2018.
[40] Know Your Meme. Pepe the Frog Meme. http://knowyourmeme.com/memes/pepe-the-frog, 2018.
[41] Know Your Meme. /pol/ KYM entry. http://knowyourmeme.com/memes/sites/pol, 2018.
[42] Know Your Meme. Rage Comics Subculture. http://knowyourmeme.com/memes/subcultures/rage-comics, 2018.
[43] Know Your Meme. Rage Guy Meme. http://knowyourmeme.com/memes/rage-guy-fffffuuuuuuuu, 2018.
[44] Know Your Meme. Rickroll Meme. http://knowyourmeme.com/memes/rickroll, 2018.
[45] Know Your Meme. Roll Safe Meme. http://knowyourmeme.com/memes/roll-safe, 2018.
[46] Know Your Meme. Russian Anti-Meme Law Meme. http://knowyourmeme.com/memes/events/russian-anti-meme-law, 2018.
[47] Know Your Meme. Smug Frog Meme. http://knowyourmeme.com/memes/smug-frog, 2018.
[48] Know Your Meme. That's The Joke Meme. http://knowyourmeme.com/memes/thats-the-joke, 2018.
[49] Know Your Meme. Trollface / Coolface / Problem? Meme. http://knowyourmeme.com/memes/trollface-coolface-problem, 2018.
[50] Know Your Meme. #TrumpAnime / Rick Wilson Controversy. http://knowyourmeme.com/memes/events/trumpanime-rick-wilson-controversy, 2018.
[51] Know Your Meme. United Kingdom Withdrawal From the European Union / Brexit Meme. http://knowyourmeme.com/memes/events/united-kingdom-withdrawal-from-the-european-union-brexit, 2018.
[52] Know Your Meme. X delivers Y at Z Meme. https://knowyourmeme.com/memes/x-delivers-y-at-z, 2018.
[53] A.-D. League. Pepe the Frog. https://www.adl.org/education/references/hate-symbols/pepe-the-frog, 2018.
[54] S. Lerner. Facebook To Ban Pepe The Frog Images Used In The Context Of Hate. http://www.techtimes.com/articles/228632/20180525/facebook-publishes-official-policy-on-pepe-the-frog.htm, 2018.
[55] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *KDD*, 2009.
[56] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. In *ICML*, 2014.
[57] S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv 1507.03228*, 2015.
[58] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack. OpenOrd: An Open-source Toolbox for Large Graph Layout. In *Visualization and Data Analysis 2011*, 2011.
[59] A. Marwick and R. Lewis. Media manipulation and disinformation online. https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf, 2017.
[60] V. Monga and B. L. Evans. Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs. *IEEE Transactions on Image Processing*, 2006.
[61] D. Olsen. How memes are being weaponized for political propaganda. https://www.salon.com/2018/02/24/how-memes-are-being-weaponized-for-political-propaganda/, 2018.
[62] Pushshift. Reddit Data. http://files.pushshift.io/reddit/, 2018.
[63] J. Ratkiewicz, M. Conover, M. R. Meiss, B. GonÃğalves, S. Patil, A. Flammini, and F. Menczer. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *Arxiv*, abs/1011.3768, 2010.
[64] Reddit. Reddit JSON Documentation. https://github.com/reddit-archive/reddit/wiki/JSON, 2018.
[65] J. Scott. *Information Warfare: The Meme is the Embryo of the Narrative Illusion*. Institute for Critical Infrastructure Technology, 2018.
[66] P. Snyder, P. Doerfler, C. Kanich, and D. McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *IMC*, 2017.
[67] O. Tsur and A. Rappoport. Don't Let Me Be #Misunderstood: Linguistically Motivated Algorithm for Predicting the Popularity of Textual Memes. In *ICWSM*, 2015.
[68] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. In *Scientific Reports*, 2012.
[69] L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting Successful Memes Using Network and Community Structure. In *ICWSM*, 2014.
[70] L. Xie, A. Natsev, J. R. Kender, M. L. Hill, and J. R. Smith. Tracking Visual Memes in Rich-Media Social Communities. In *ICWSM*, 2011.
[71] I. Yoon. Why is it not just a joke? Analysis of Internet memes associated with racism and hidden ideology of colorblindness. *Journal of Cultural Research in Art Education*, 33, 2016.
[72] S. Zannettou, B. Bradlyn, E. De Cristofaro, M. Sirivianos, G. Stringhini, H. Kwak, and J. Blackburn. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? In *WWW Companion*, 2018.
[73] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the Origins of Memes by Means of Fringe Web Communities (Extended Version). *arXiv preprint arXiv:1805.12512*, 2018.
[74] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *IMC*, 2017.
[75] C. Zauner, M. Steinebach, and E. Hermann. Rihamark: perceptual image hash benchmarking. In *Media Forensics and Security*, 2011.

A CLUSTERING PARAMETER SELECTION

Our implementation uses the DBSCAN algorithm with a clustering threshold equal to 8. To select this threshold, we perform the clustering step while varying the distances. Table 6 shows the number of clusters and the percentage of images that are regarded as noise by the clustering algorithm for varying distances. We observe that for distances 2-4 we have a substantially larger percentage of noise, while with distance 10 we have the least percentage of noise. With distances between 6 and 8 we observe that we get a larger number

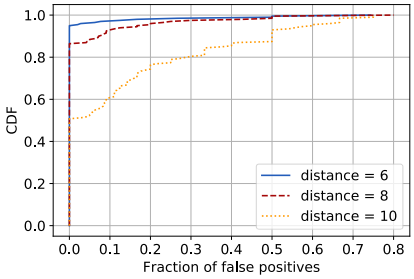


Figure 13: Fraction of false positives in clusters with varying clustering distance.

Distance	#Clusters	%Noise
2	30,327	82.9%
4	34,146	78.5%
6	37,292	73.0%
8	38,851	62.8%
10	30,737	27.8%

Table 6: Number of clusters and percentage of noise for varying clustering distances.

of clusters than the other distances, while the noise percentages are 73% and 63%, respectively. To further evaluate the clustering performance for varying distances, we randomly select 200 clusters and manually calculate the number of images that are false positives within each cluster. Fig. 13 shows the CDF of the false positive fraction in the random sample of clusters for distances 6,8, and 10 (we disregard distances 2-4 due to the high percentage of noise). We observe that with distance 10 we have an unacceptably high fraction of false positives, while for distances 6-8 we observe a similar behavior with slightly more false positives with distance 8 (in both cases the overall false positives are below 3%). A question that arises here is what is the impact of those false positives in the overall dataset? To shed light to this question, we find all posts that contain false positives and true positives (from the random sample of 200 clusters with distance 8). We find that the false positives have little impact as they occur substantially fewer times than true positives: the percentage of true positives over the set of false positives and true positives is 99.4%. Therefore, due to the larger number of clusters, the acceptable false positive performance, and the smaller percentage of noise (when compared to distances 2-6), we elect to use as a threshold the perceptual distance that is equal to 8.

B KYM AND CLUSTERING ANNOTATION EVALUATION

While KYM might not be a house hold name, the site is seemingly the largest curated collection of memes on the Web, i.e., KYM is as close to an "authority" on memes as there is. That said, crowdsourcing *is* an aspect of how KYM works, and thus there are questions as to how "legitimate" some of the content is. To this end, we set out to measure the quality of KYM by sampling a number of pages and manually examining them. This is clearly a subjective task, and a fully specified definition of what makes a valid meme is approximately as difficult as defining "art." Nevertheless, the authors of this paper have, for better or worse, collectively spent thousands of hours immersed in the communities we explore, and thus, while we are not confident in providing a strict definition of a meme, we are confident in claiming that we know a meme when we see it.

Using the same randomly selected 200 clusters as mentioned in Appendix A, we visited each KYM page the cluster was tagged with and noted whether or not it properly documented what we consider an "actual" meme. The 200 clusters were mapped to 162 unique KYM pages, and of these 162 pages, 3 (1.85%) we decided were "bad." This is mainly due to the lack of completeness and relatively high number of random images in the gallery (see [38, 52] for some examples of "bad" KYM entries).

Next, we set out to determine whether the label (i.e., KYM page) assigned to each of our randomly sampled clusters was appropriate. Using three annotators, for each cluster we examined the KYM page, the medoid of the cluster, and the images in the cluster itself and noted whether the label does in fact apply to the cluster. Here, again, there is a great degree of subjectivity. To reign some of the subjectivity in, we used the following guidelines:

- If the exact image(s) in the cluster appear in the KYM gallery, then the label is correct.
- For images that do not appear in the KYM gallery, if the label is *appropriate*, then it is a correct labeling.

There are some important caveats with these guidelines. First, KYM galleries are crowdsourced, and while curated to some extent, the possibility for what amounts to random images in a gallery *does* exist; however, based on our assessment of KYM page validity, this occurs with low probability. Second, we considered a label correct if it was *appropriate*, even if it was not necessarily the *best* possible label. For example, as our results show, many memes are related, and many images mix and match pieces of various memes. While it is definitely true that there might be better labels that exist for a given cluster, this straightforward and comprehensible labeling process is sufficient for our purposes. We leave a more in-depth study of the subjective nature of memes for future work. Finally, it is important to note that memes are a *cultural* phenomenon, thus the potential for cultural bias in our annotation is possible. To that end, we note that our annotators were born in three different countries (USA, Italy, and Cyprus), only one is a native English speaker, and two have spent substantial time in the US.

After annotating clusters, we compute the Fleis agreement score (κ). With our cluster samples, we achieve a κ of 0.67, which is considered "substantial" agreement. Finally, for each cluster we obtain the majority agreement of all annotators to assess the accuracy of our annotation process; we find that 89% of the clusters had a legitimate annotation to a specific KYM entry.