

# Detecting Spammers and Content Promoters in Online Video Social Networks

Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida,  
Jussara Almeida and Marcos Gonçalves

Federal University of Minas Gerais - Brazil

# Motivation

- Video is a trend on the Web
  - video forum, video blog, video advertises, political debates
  - 77% of the U.S. Internet audience viewed online videos
- Explosion of user generated content
  - YouTube has 10 hours of videos uploaded every minute

**User generated videos are susceptible to various opportunistic user actions**


# Example of Video Spam

**You Tube** Broadcast Yourself™ Global (Todos) Português Inscreva-se Lista rápida (0) Ajuda Fazer login

Página inicial Vídeos Canais Comunidade




Vídeos Pesquisar avançado Enviar

### Polska-Czechy 2:1 wszystkie bramki

 **Polska-Czechy 2:1 wszystkie bramki**  
03:53  
:D POLSKA-CZECHY 2:1 W ELIMINACJACH DO MS 2010 NA  
MAGICZNYM STADIONIE W CHORZOWIE :D THX LEO


De: [Kran6](#)  
Data de entrada: 2 meses  
atrás  
Vídeos: 8

**Respostas ao video** (9 respostas) Reproduzir todas as respostas ao video




**Pornography**

De: moppix  
Exibições: 278394  
Resposta: 9  
02:03 ★★★★★




**Juninho two new amazing free kic...**

De: brazilabras...  
Exibições: 41033  
Resposta: 7  
01:11 ★★★★★



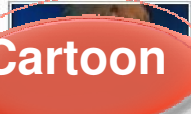

**6 Years Old Kid Amazing Football...**

De: yhnell18  
Exibições: 73  
Resposta: 6  
04:15 ★★★★★




**Amazing Football...**

De: yhnell18  
Exibições: 2128  
Resposta: 4  
00:47 sem avaliação




**Cartoon**

De: luisianaluiza  
Exibições: 3976  
Resposta: 3  
03:46 ★★★★★



**Street Panna Talents Part 1 Trailer**

De: StreetPannaT...  
Exibições: 3318  
Resposta: 2  
00:36 ★★★★★



**Polska-Czechy 2:1 Chorzów 2008 E...**

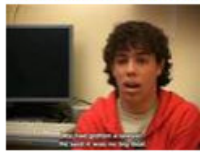
De: mojasokolka  
Exibições: 19553  
Resposta: 1  
02:32 ★★★★★

+ Google Vídeos Pesquisar

# Example of Promotion



## Eric and the Army of the Phoenix (1/5)



Eric and the Army of the Phoenix (1/5)  
9:48

An incredible but true story: Spanish authorities prosecute child for terrorism when he e-mails companies requesting labelling in Catalan language, using Phoenix monicker from Harry Potter books. Poli ([more](#))



From: ericielfenix  
Joined: 2 years ago  
Videos: 6

## Video Responses (8352 Responses)

[Play All Video Responses](#)



Torroella de Montgrí (Baix Empordà)  
0.58K habitar  
0:05  
160 views  
danimorph  
★★★★★



Torrent (Baix Empordà)  
190 habitar  
0:05  
22 views  
danimorph  
no rating



Tallada d'Empordà (Baix Empordà)  
367 habitar  
0:05  
27 views  
danimorph  
no rating



Serra de Daró (Baix Empordà)  
189 habitar  
0:05  
36 views  
danimorph  
no rating



Santa Cristina d'Aro (Baix Empordà)  
4.21K habitar  
0:05  
111 views  
danimorph  
no rating



Sant Feliu de Guíxols (Baix Empo...)  
20.8K habitar  
0:05  
101 views  
danimorph  
★★★★★



Rupia (Baix Empordà)  
213 habitar  
0:05  
67 views  
danimorph  
no rating



Regencós (Baix Empordà)  
304 habitar  
0:05  
63 views  
danimorph  
no rating



la Pera (Baix Empordà)  
438 habitar  
0:05  
27 views  
danimorph  
no rating



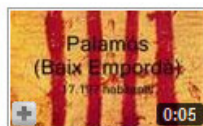
Parlavà (Baix Empordà)  
352 habitar  
0:05  
53 views  
danimorph  
no rating



Pals (Baix Empordà)  
2.40K habitar  
0:05  
40 views  
danimorph  
no rating



Palau-sator (Baix Empordà)  
280 habitar  
0:05  
70 views  
danimorph  
no rating



Palamós (Baix Empordà)  
17.1K habitar  
0:05



Palafrugell (Baix Empordà)  
21.3K habitar  
0:05



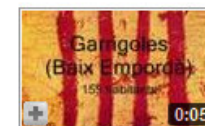
Mont-ras (Baix Empordà)  
1.85K habitar  
0:05



Jafre (Baix Empordà)  
381 habitar  
0:05



Gualta (Baix Empordà)  
358 habitar  
0:05



Garrigoles (Baix Empordà)  
155 habitar  
0:05

# Negative Impact of Promotion and Spam

- Challenges for users in identifying video promotion and spam
  - consumes system resources, especially bandwidth
  - compromise user patience and satisfaction with the system
- Pollution in top lists
- Difficulty in ranking and recommendation
  - Promoted or spam videos may be temporarily ranked high

# Goal

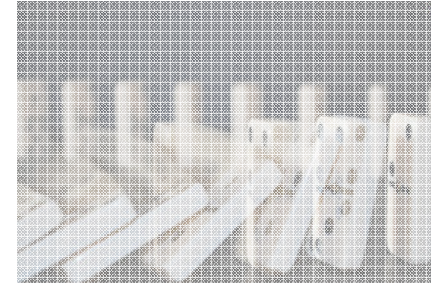
- **Detect video spammers and promoters**
- 4-step approach
  1. Sample YouTube video responses and users
  2. Manually create a user test collection  
(promoters, spammers, and legitimate users)
  3. Identify attributes that can distinguish spammers and promoters from legitimate users
  4. Classification approach to detect spammers and promoters



**Part1.  
Motivation  
& Problem**



**Part2.  
4-step  
approach**



**Part3.  
Experimental  
results**

# Step1. Sampling video responses



- **Approach:** Collect entire weakly connected components
  - Follow both directions: video responses and video responded
  - Collect all videos of each user found
  - This approach allow us to use several social network metrics
- Collected 701,950 video responses and 381,616 video topics, 264,460 users in 7 days in January, 2008

# Step2. Create Test Collection

## Desired Properties

- 1) Have a significant number of users in each class
- 2) Include spammers and promoters which are aggressive in their strategies
- 3) Include a large number of legitimate users with different behavioral profiles

# Step2. Create Test Collection

- Users selected according to three strategies
  - 1) Manually identified 150 suspect in the top 100 most responded lists
  - 2) Randomly select 300 users from those who posted video responses to videos in the top 100 most responded lists
  - 3) Collected 400 users across 4 different levels of interaction
    - sent and received video responses
- Volunteers analyze users and videos
  - Conservative approach -> favor legitimate
  - Agreement in 97% of the analyzed videos

**TOTAL: 829 users, 641 legitimate, 157 spammers, 31 promoters**

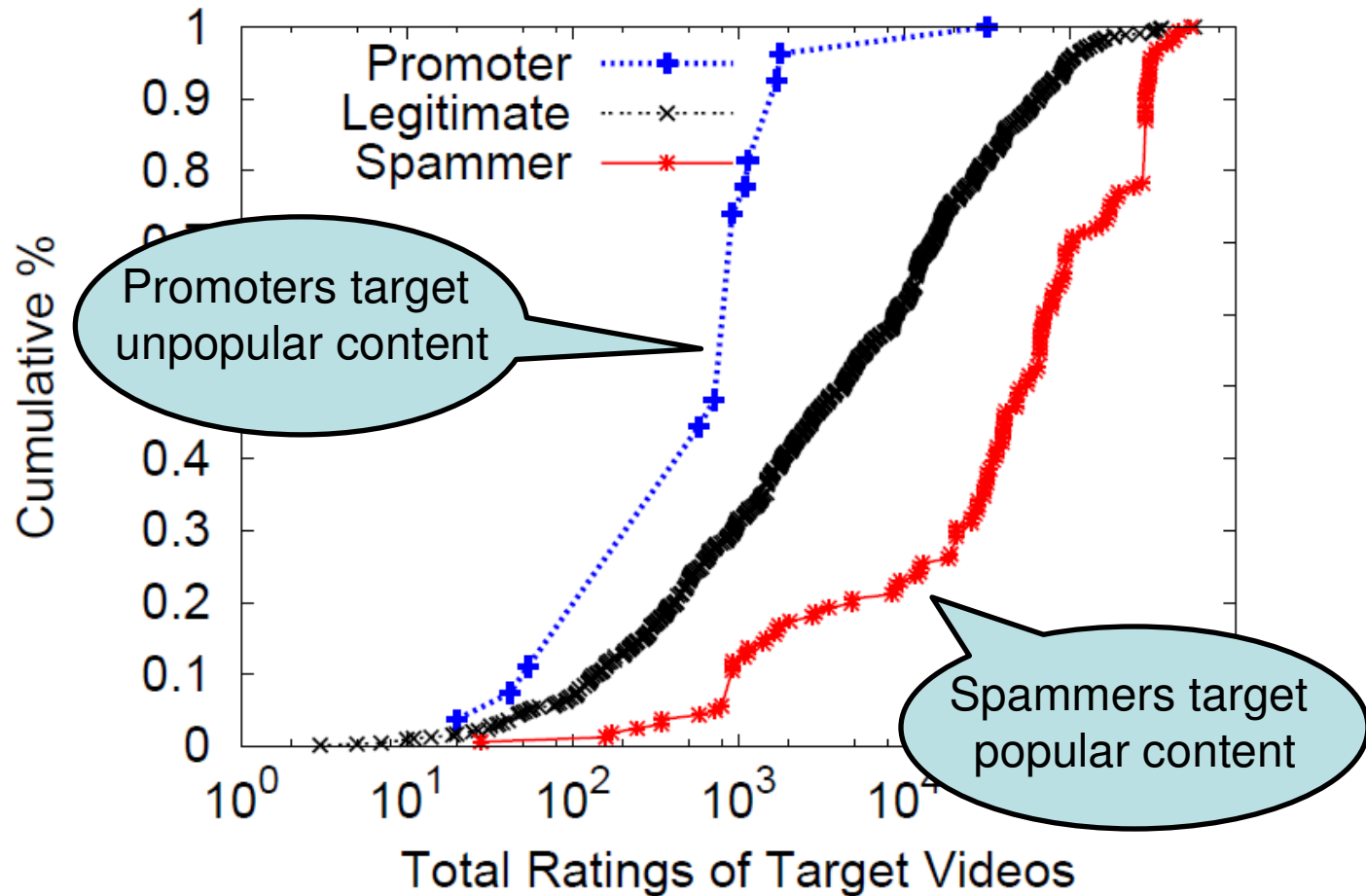
# Step3. Attributes

- **User-Based:**
  - number of friends, number of subscriptions and subscribers, etc
- **Video-Based:**
  - duration, numbers of views and of comments received, ratings, etc
- **Social Network:**
  - clustering coefficient, betweenness, reciprocity, UserRank, etc

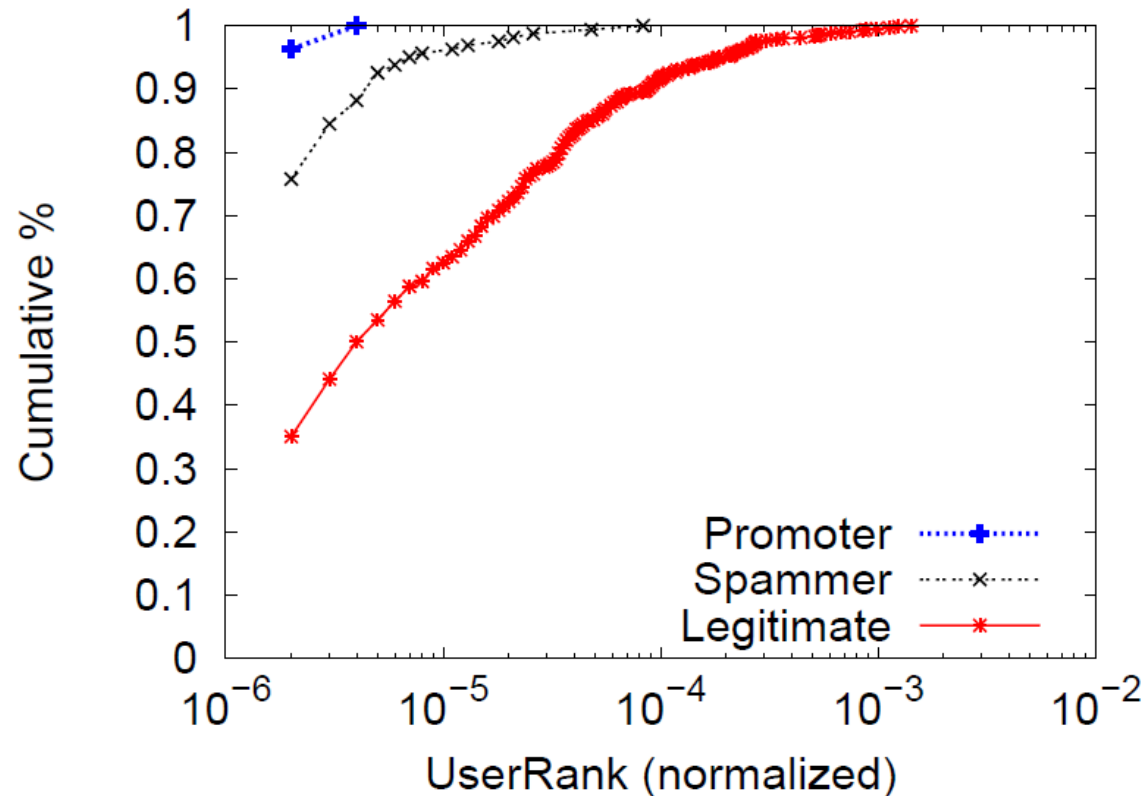
## Feature Selection: $\chi^2$ ranking

Attribute Set	Top 10	Top 20	Top 30	Top 40	Top 50
Video	9	18	25	30	36
User	1	2	4	7	9
SN	0	0	1	3	5

# Distinguishing classes of users (1)



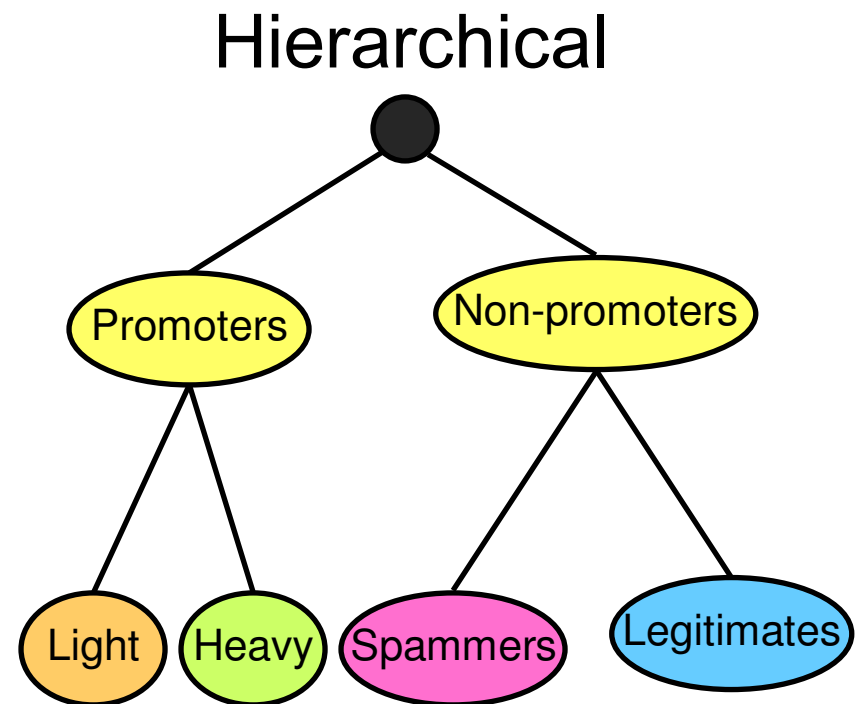
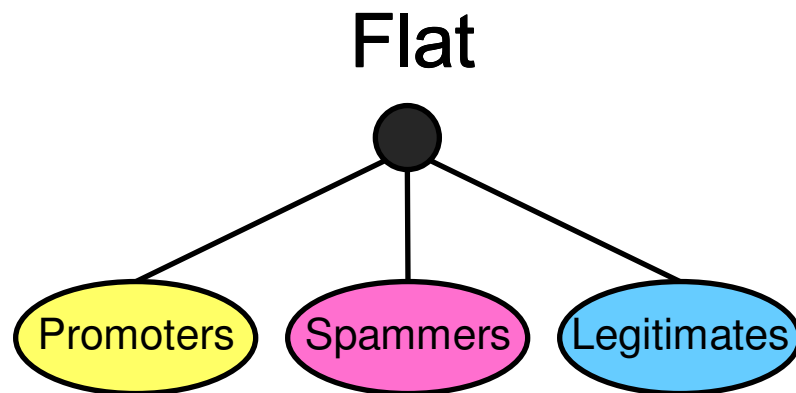
# Distinguishing classes of users (2)



Even low-ranked features have potential  
to separate classes apart

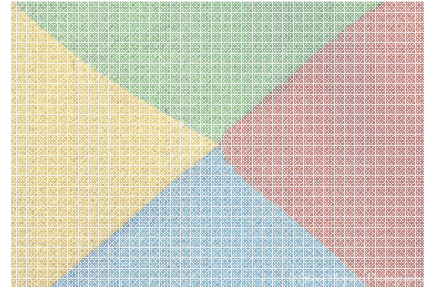
# Step4. Classification Approach

- SVM (Support vector machine) as classifier
  - Use all attributes
  - Two classification approaches





**Part1.  
Motivation  
& Problem**

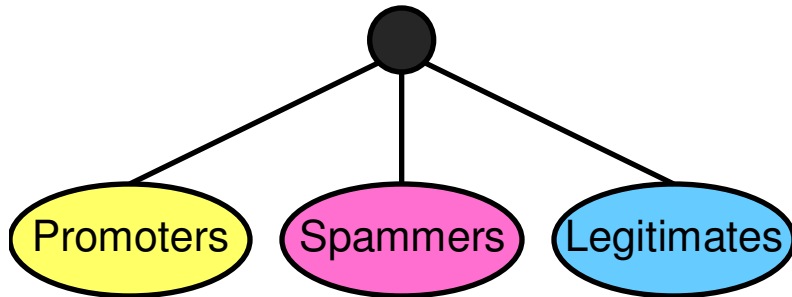


**Part2.  
4-step  
approach**



**Part3.  
Experimental  
results**

# Flat Classification

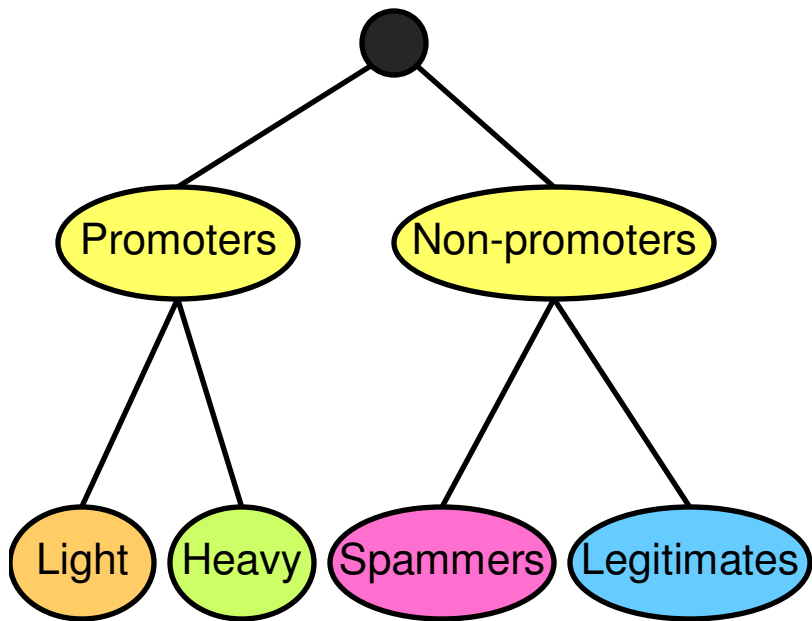


- Correctly identify majority of promoters, misclassifying a small fraction of legitimate users.
- Detect a significant fraction of spammers but they are much harder to distinguish from legitimate users.
  - Dual behavior of some spammers

		Predicted		
		Promoter	Spammer	Legitimate
True	Promoter	96.13%	3.87%	0.00%
	Spammer	1.40%	56.69%	41.91%
	Legitimate	0.31%	5.02%	94.66%

- Micro F1 = 88% (predict the correct class 88% of cases)

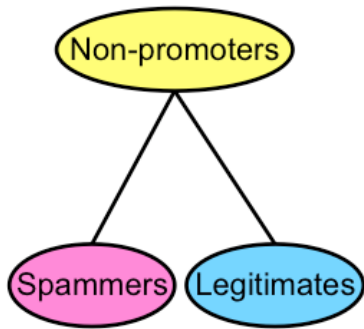
# Hierarchical Classification



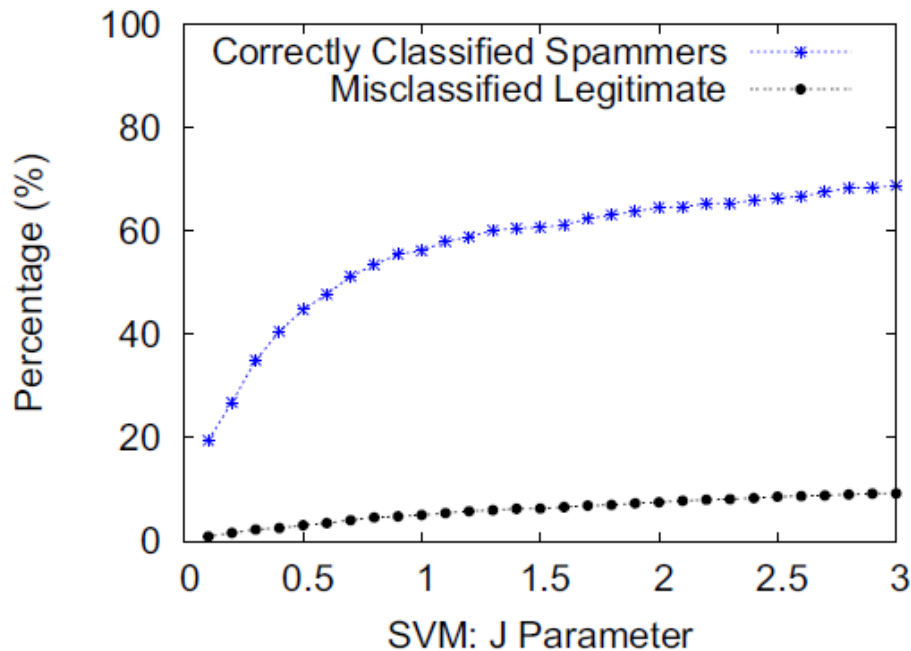
- **Goal:** provide flexibility in classification accuracy
- **First Level:**
  - Most promoters are correctly classified
  - Statistically indistinguishable compared with flat strategy

		Predicted	
		Promoter	Non-Promoter
True	Promoter	92.26%	7.74%
	Non-Promoter	0.55%	99.45%

# Distinguishing Spammers from Legitimate users



		Predicted	
		Legitimate	Spammer
True	Legitimate	95.09%	4.91%
	Spammer	41.27%	58.73%

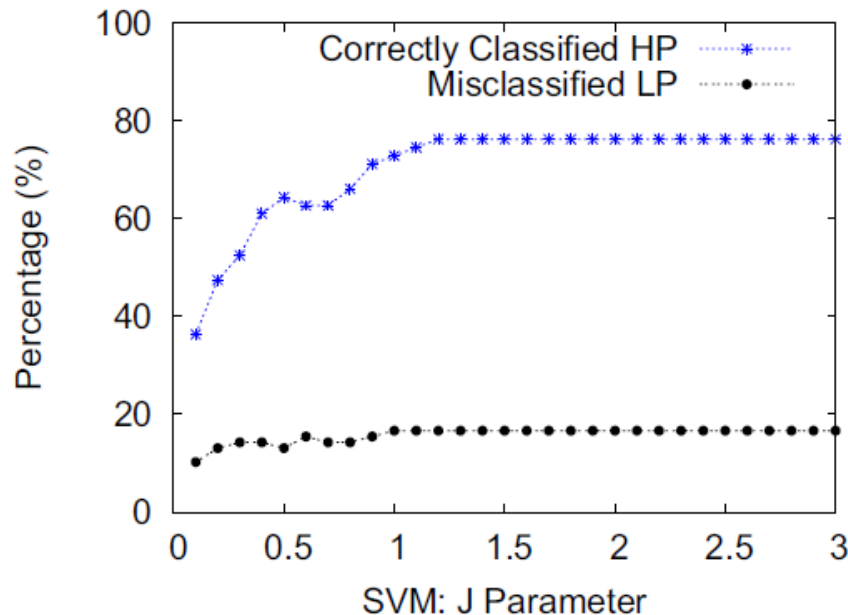
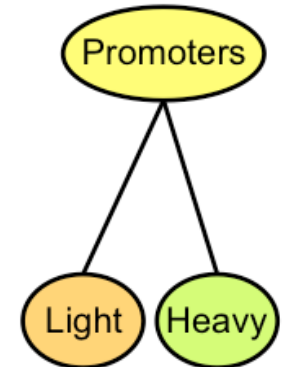


- **J = 0.1**: correctly classify 24% spammers, misclassifying <1% legitimate users
- **J = 3**: correctly classify 71% spammers, paying the cost of misclassifying 9% legitimate users

# Distinguishing Promoters

- Heavy promoters could reach the top-100 in one day
- Light promoters associated with a collusion attack

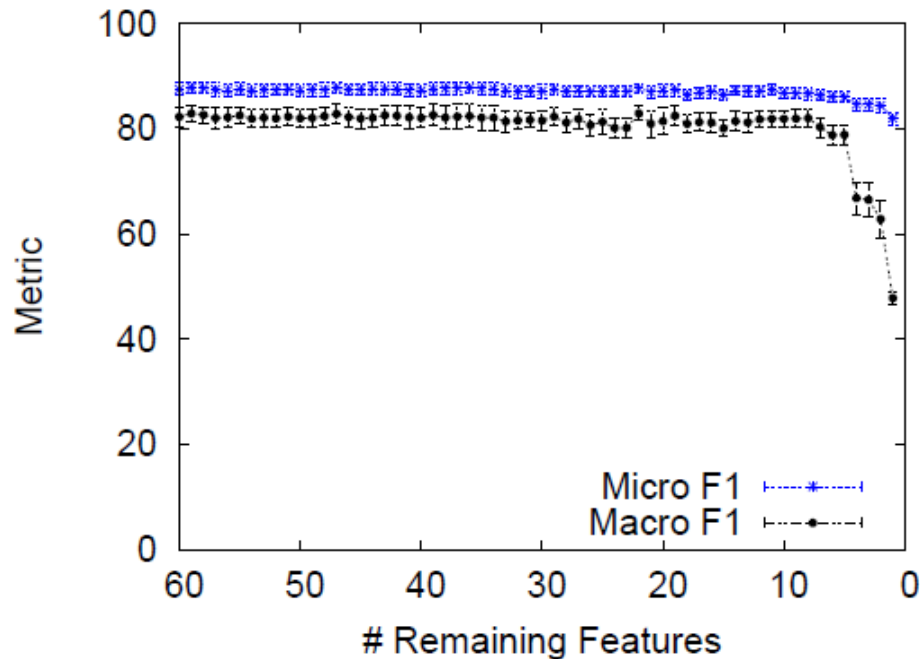
		Predicted	
		Light Promoter	Heavy Promoter
True	Light Promoter	83.33%	16.67%
	Heavy Promoter	27.12%	72.88%



- $J = 0.1$ : correctly classify 36% of heavy promoters at the cost of misclassifying 10% of light promoters
- $J = 1.2$ : correctly classify 76% of heavy promoters at the cost of misclassifying 17% light ones

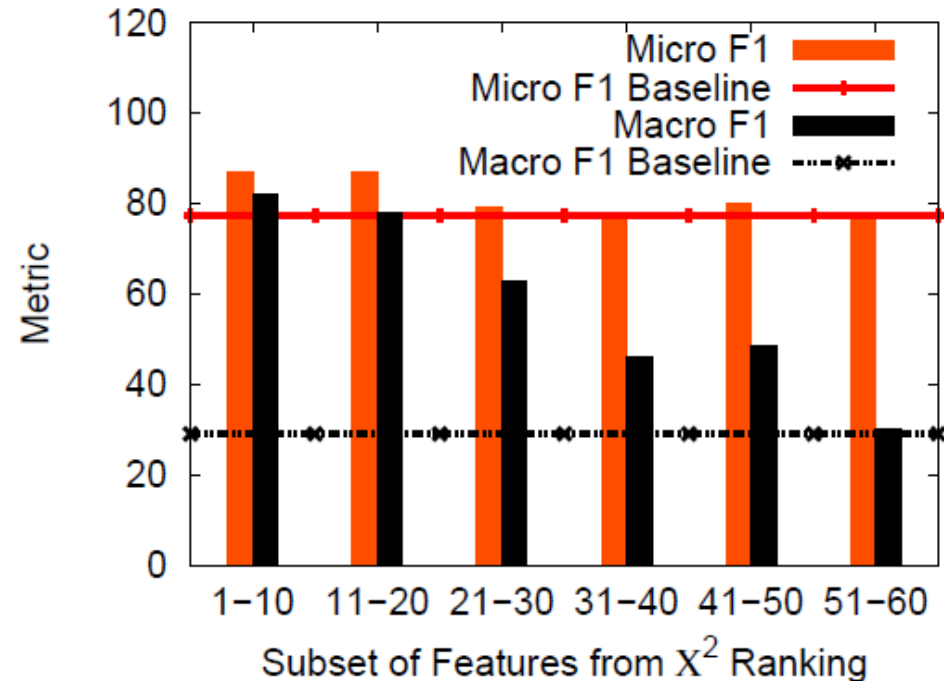
# Reducing the Attribute Set

## Scenario 1



Classification approach is effective even with a smaller, less expensive set of attributes

## Scenario 2



Different subsets of features can obtain competitive results

# Conclusions

- First approach to detect spammers and promoters
  - Attribute identification
  - Creation of a test collection
    - available at [www.dcc.ufmg.br/~fabricio](http://www.dcc.ufmg.br/~fabricio)
  - Classification approach
    - Correctly identify majority of promoters
    - Spammers showed to be much harder to distinguish
      - trade-off between detect more spammers at the cost of misclassifying more legitimate users