Computer Networks xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Talent scouting in P2P networks

N. Koenigstein *, Y. Shavitt

School of Electrical Engineering, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Spatial data mining
P2P networks
Information retrieval

ABSTRACT

Record labels would like to identify potential artists as early as possible in their career, before other companies approach the artists with competing contracts. However, there is a huge number of new artists, and the process of identifying the ones with high success potential is labor intensive. This paper demonstrates how data mining in P2P networks can be used together with social marketing theories in order to mechanize most of this detection process.

Using a unique intercepting system over the Gnutella network we captured an unprecedented amount of geographically identified queries, allowing us to investigate the diffusion of music related content in time and space. Our solution is based on the observation that successful artists, start by growing a discernible stronghold of fans in their hometown area, where they are able to perform and market their music. Only then they manage to breakthrough to national fame. In a file sharing network, their initial local success is reflected as a delta function spatial distribution of content queries. Using this observation, we devised a detection algorithm for emerging artists that suggests a short list of artists with breakthrough potential, from which we showed that about 30% translate the potential to national success.

1. Introduction

Record label companies are constantly looking for the "next big thing". Each year, a small number of new artists succeed in breaking out of their anonymity and rise to stardom. The artists and repertoire (A&R) divisions in record labels are responsible for discovering these artists out of the masses of unknown talent on the market. A&R executives rely mostly on the word of mouth of trusted associates, critics and business contacts. Human scouts are expected to understand the current tastes of the market, but in the case of unfamiliar new artists, they have little else to rely on but their gut instinct. They therefore, tend to favor artist coming from their own city, where they can "feel the scene" [1]. Nonetheless, their predictions are usually wrong, and only around 20% of signed artists will make money for their label.[1] Locating artists with high success potential is thus of great importance for the music industry.

We examine a database of query strings to the Gnutella file sharing network, and use our understanding of the Gnutella protocol in order to identify the ones which can be located geographically. Since our aim is to gain advantage in the artists scouting market, we look at fairly rare queries, those that are not even on the top 2000 list, trying to detect emerging artists with higher potential to make a national level breakthrough (in the US). We mathematically model the popularity diffusion patterns of emerging artists in their first steps, and explain why local popularity is an important factor when new, previously unknown artists are considered. This research demonstrates how P2P query strings can be used by the music industry for intelligent decision making.

* Corresponding author. Tel.: +972 522653599.
E-mail addresses: noamk@eng.tau.ac.il (N. Koenigstein), shavitt@eng.tau.ac.il (Y. Shavitt).

[1] Taken from: www.telegraph.co.uk/technology/3778304/Could-software-find-the-X-Factor.html (Accessed: June 2010).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.10.021

The rest of this paper is organized as follows: In Section 2, we explain how the database of Gnutella geo-aware query strings is created. Then, in Section 3, we review the small world model for the spatial and temporal diffusion of new innovations. In Section 4 we discuss the diffusion of digital content from emerging artists. Finally, in Section 5 we describe the detection algorithm and test its predictions in Section 6.

2. Data collection

Our scouting algorithm (described in Section 5), can be applied in different P2P networks, as well as music related websites such as MySpace or even YouTube. In this study we used data collected from the Gnutella network which was the most popular file sharing network on the Internet at the time of data collection [2].

Capturing a large quantity of Gnutella queries is achieved by deploying several hundred ultrapeer nodes[2] in the network. This has been done in several previous academic studies [3–6]. However, inferring the origin IP address (essential for the geographical mapping), is not a trivial task. The Gnutella protocol does not maintain an origin address in queries. Instead it keeps an "Out Of Band" (OOB) return IP address, which is the origin IP address in principle. However, for firewalled clients which cannot accept connections from the outside the OOB address belongs to the ultrapeer acting as a proxy on behalf of the query originator. For such clients there are no known methods to determine their IP address. Furthermore, there is no explicit indication in the query whether it was issued from a firewalled client or not. In this study we have used the same data collection system as in [6]. More details about the Gnutella protocol, and the system implementations can be found in their paper. Here we will only explain how we overcome the above problem of IP resolution, as this is crucial for the geographical mapping of the queries used by our algorithm.

To overcome the above difficulties, we wish to use only queries originating from non-firewalled clients where the OOB field carries the leaf's IP address. We therefore, devised a process to distinguish firewalled from non-firewalled queries. Our technique is based on another field carried by the query, the hop count, which is set to 1 by the originator, and incremented by one each time a node forwards the query to its neighbors. To understand this technique, observe the small network in Fig. 1. The figure depicts an intercepting node along with other ultrapeers and leafs. Ultrapeer B is directly connected to the intercepting node. Thus, any query that traversed only a single hop must have come from it. Leaf A, leaf C (firewalled), and ultrapeer D are at a distance of two hops away. All queries coming from A, C and D, will have a hop count of two. Queries from Leaf A and Ultrapeer D will contain their own IP address in the OOB field. However, queries originating at C will contain B's OOB return address as C is firewalled or otherwise unable to accept incoming connections. As we are directly connected to ultrapeer B, we can simply compare the query's OOB address with B's address. If they are not identical, the query must have come from A or D, and the address is guarantied to be the origin's address. If the query contain's B's address but passed two hops, it must be coming from a firewalled client (leaf C) using ultrapeer B as a proxy. In this case C's address is not available, and the query is not recorded. Ultrapeer F and leaves E (firewalled) and G are at a distance of three hops or more. When we intercept their queries we cannot know whether the OOB IP address belongs to them, or perhaps to ultrapeers D or F (acting as a proxy). Thus, any query that traversed three hops or more is discarded. As a result, an intercepting node records traffic originating from its immediate neighborhood only (having a hop count $\leqslant 2$), thus requiring a massive deployment of such nodes.

The described setting eliminates most of the bias against popular queries which travel only short distances before being satisfied. Discarding such queries cancels the advantage of "rare" queries that stay in the network longer. However, this setting does introduce a bias against queries from firewalled clients, as only queries that can receive incoming connections are recorded.

2.1. Dataset statistics

The removal of queries which traveled more than two hops, non-Limrewire clients, firewalled queries and non-OOB enabled queries, amounts for approximately 75% of the intercepted queries. We remained with 25–40 million IP identified queries every day. We then used the commercial IPligence database to resolve the geographical location of the IP addresses bounding country, state, city, latitude and longitude to each query string. This allowed us to pin point the source of each query string to the level of cities and sometimes even smaller areas like the boroughs of NYC. Since we concentrate our study on American artists, we removed all the non-US queries reducing an additional 55–60% of the data records.

Our data-set comprised of query strings collected over a period of nine and a half months from mid October 2006 until July 2007. The activity on the Gnutella networks increases by 20–25% over the weekend [6]. We thus used weekly samples taken on a Saturday or a Sunday of every week of that period. The sample from the 51st week of 2006 and the samples from the 24th and 25th weeks of 2007 were not recorded as a result of technical difficulties. We thus remained with 38 samples instead of 41. The total number of unique geo-aware query strings processed in this study is **310,380,190**, making it the largest P2P queries study thus far.

Using the geo-aware query strings we generated weekly global (national) and local (per city) popularity charts. For each string, its global and local popularity was calculated by aggregating the number of appearances intercepted. The chart rankings were calculated by sorting the queries according to popularity. This means that in principle, more than a single string can be ranked in a given position. However, since we dealt with millions of queries this rarely ever happened among the popular strings. These popular-

[2] Ultrapeer nodes are nodes that were selected to form the backbone of the Gnutella network. As such they route queries and responses for other nodes connected to them.

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

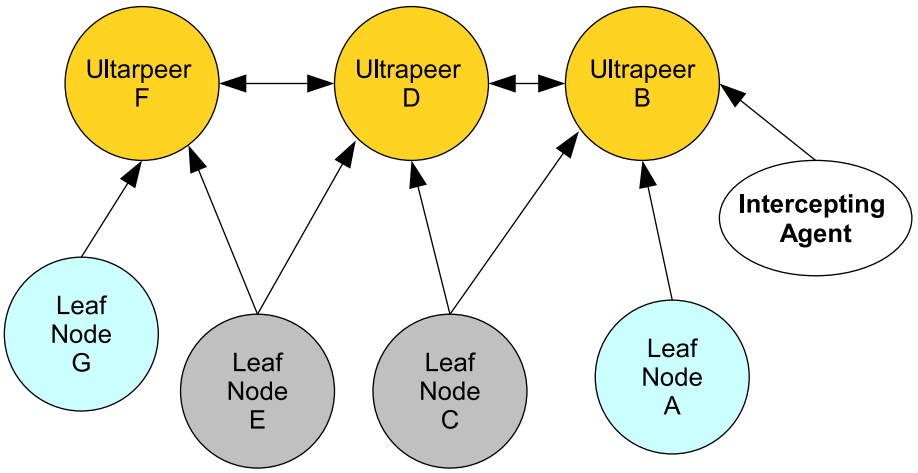


Fig. 1. Two-tier Gnutella segment.

ity charts and the Billboard charts were shown to have high correlation [7].

Queries content classification: Manually classifying the top 500 most popular queries, we found that 68.11% of the files were music related while 22.01% were adult content. These two categories dominate the Gnutella network accounting together for 90.12% of the queries.

3. Spatio-temporal properties of content diffusion

In order to understand emerging artists diffusion in P2P networks, let us first review the small world model for the spatial and temporal diffusion of new products. Small-world modeling assumes that the main driver behind a product's growth is communication between individuals. A successful product is noticeable by the formation of adopter clusters around early adopters. These clusters represents areas where the new product is spreading as a result of the *Word Of Mouth* (WOM) effect and can be used in order to predict a product's future success.

3.1. The word of mouth effect

In his groundbreaking work from 1973 "The Strength of Weak Ties" [8], sociologist Granovetter suggested an explanation of how micro-level interactions between individuals affects macro level phenomena. Relationships between individuals can be modeled as a network, where persons are nodes connected through ties to other individuals. The interpersonal ties that connect individuals are assigned a "strength" parameter that is a function of the amount of time, emotional intensity, intimacy, and reciprocal services which characterize the tie. Using the strength property, we can label ties as "weak" or "strong". For example, a tie between two siblings can be considered a strong tie, while a tie between two persons that chat on a random meeting while waiting in a bus station can be considered a weak tie. This modeling can be used to understand the way information is spread in a community; a phenomenon known as the WOM effect. Adopting a consumer research orientation, this effect was investigated in [9] where it was found that while strong ties were more influential in consumer decisions at the micro level, the weak ties were more important in facilitating "bridges" that allow information to travel between different social subgroups at the macro level. Small world networks mathematically model Granovetter's ideas of strong and weak ties and the WOM effect [10]. Garber et al. used it to investigate a new product spatial diffusion after it was introduced to the market [11].

The small-world model depicts the market as a binary matrix, the elements of which represent individuals in different locations. A '1' represents a consumer who bought the product (adopted) and a '0' is a consumer who has not (a non-adopter). Each consumer can interact with his acquaintances and influence them to purchase the new product. A persons's group of acquaintances consists mostly of people in his close vicinity (neighbors) and a small group of people outside his vicinity. Similarly each cell in the matrix can interact with its neighboring cells up to some specified range and a small number of random cells outside the cell's vicinity. In a classical small-world model the proportion of distant links is 5% at most. Beyond this level, the social system becomes similar to a random network [12].

There are two types of events that cause a non-adopter to buy the product:

- *Internal factors:* An interaction with an acquaintance adopter that influence the consumer to buy the product (The WOM effect). Such an event happens with probability q due to either the person's strong (close) or weak connections.
- *External factors*: An individual decides to adopt because of external influence such as advertising. This event happens with probability p.

Therefore the probability that a non-adopter will adopt at a time slot t is:

$$prob(t) = 1 - (1-p)(1-q)^{\nu(t)+r(t)} \quad (1)$$

where $\nu(t)$ is the number of previous adopters with whom the non-adopter have connections in his vicinity (strong

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

ties) and $r(t)$ is the number of adopters with whom he has ties outside his vicinity (weak ties).

The adoption pattern of the new product depends on the numeric values of p and q. High values of p mean effective advertisement which cause a uniformly distributed increase in new adopters. However a product's success depends mostly on the value of q, which models the WOM effect. High values of q is an indication that people "like" the new product, and so adopters effect non-adopters to purchase it. Low values of q mean adopters are not satisfied with the product and the product is likely to fail. Even an excellent advertising campaign can not salvage a product that disappoints consumers. Therefore in small-world modeling we can predict a product's success according to the values of q, which represents the WOM effect and social imitation [11].

3.2. The divergence measure as an indicator of innovation success

High values of q causes the formation of adopter clusters at the early stages of a new product spatial diffusion. These adopter clusters represents areas where the new product is spreading as a result of the WOM effect. When q is low, the product diffuses uniformly in space due to the "external" marketing efforts. Looking at a new product spatial sales distribution, one should be able to distinguish between the uniform distribution and the presence of adopter clusters.

Garber et al. [11] suggested the Kullback–Leibler (KL) divergence measure to predict products success probability. The KL divergence for the difference between two probability vectors P and Q is defined as follows:

$$D_{KL}(P\|Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{2}$$

If we take P to be the product's sales distribution vector, where component $P(i)$ is the relative amount of sales occurred in region i, and Q is the uniform distribution, then we receive a numeric value that measures how much the sales distribution differs from the uniform distribution. The minimum value, zero, is obtained when sales are uniformly distributed.

It is advisable to use P as the product sales distribution and Q as the uniform distribution (and not vice versa). Using the limit: $\lim_{x\to 0} x\log x = 0$, will avoid division by zero, in the case that in one of the regions there were no sales at all.

The KL divergence is non-negative but asymmetric and not always finite. Instead the Jensen–Shannon (JS) divergence was suggested:

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \tag{3}$$

where $M = \frac{1}{2}(P+Q)$

When a new product becomes popular (after being adopted by many consumers), the spatial distribution of buyers becomes more similar to the uniform distribution resulting in a decline of the divergence measure [11]. Eventually there are adopters everywhere, and no remains of the initial clusters. At that stage both the distribution vector of a successful product and that of a failure are close to uniform. They differ, of course, in the total number of adopters, which is what determines if the product succeeded or failed. In the next section, we show a very similar behavior with songs that belong to emerging artists.

3.3. δ distribution of adopters

If Q is fixed, the distribution P that maximize the divergence $D(P\|Q)$ is the δ distribution (i.e., $P = [0,\ldots,0,1,0,\ldots,0]$). This is true for both the Kullback–Leibler divergence and the Jensen–Shannon divergence, and for any fixed distribution Q. In order to show this, we first need to prove that $D_{kl}(P\|Q)$ and $D_{js}(P\|Q)$ are convex in P.

Theorem 3.1. *$D_{kl}(P\|Q)$ is convex in the pair (P,Q) i.e., if (P_1,Q_1) and (P_2,Q_2) are two pairs of probability mass distribution, then*

$$D_{kl}(\lambda P_1 + (1-\lambda)P_2\|\lambda Q_1 + (1-\lambda)Q_2) \leqslant \lambda D_{kl}(P_1\|Q_1) + (1-\lambda)D_{kl}(P_2\|Q_2)$$

for all $0 \leqslant \lambda \leqslant 1$

Proof. See Theorem 2.7.2 in [13] □

Corollary 3.2. *For any fixed probability mass distribution Q, $D_{kl}(P\|Q)$ is convex in P.*

Proof. If we set $Q_1 = Q_2 = Q$, then by using Theorem 3.1 we get:

$$D_{kl}(\lambda P_1 + (1-\lambda)P_2\|Q) \leqslant \lambda D_{kl}(P_1\|Q) + (1-\lambda)D_{kl}(P_2\|Q)$$ □

So far, we established the convexity of $D_{kl}(P\|Q)$ in P. Let us show the same property for $D_{js}(P\|Q)$:

Lemma 3.3. *For a fixed value of Q and $M = \frac{1}{2}(P+Q)$, $D_{kl}(P\|M)$ is convex in P.*

Proof

$$\begin{aligned} &D_{kl}\left(\lambda P_1 + (1-\lambda)P_2\|\frac{\lambda P_1 + (1-\lambda)P_2}{2} + \frac{Q}{2}\right) \\ &= D_{kl}\left(\lambda P_1 + (1-\lambda)P_2\|\lambda\left(\frac{P_1}{2}+\frac{Q}{2}\right) + (1-\lambda)\left(\frac{P_2}{2}+\frac{Q}{2}\right)\right) \\ &\leqslant \lambda D_{kl}\left(P_1\|\frac{P_1}{2}+\frac{Q}{2}\right) + (1-\lambda)D_{kl}\left(P_2\|\frac{P_2}{2}+\frac{Q}{2}\right) \end{aligned} \tag{4}$$

where in (4) we used the convexity property of $D_{kl}(P\|Q)$ in the pair (P,Q) (Theorem 3.1). □

Lemma 3.4. *For a fixed value of Q and $M = \frac{1}{2}(P+Q)$, $D_{kl}(Q\|M)$ is convex in P.*

Proof. The proof here is very similar (but not identical) to the proof in Lemma 3.3. □

Corollary 3.5. *For a fixed Q, $D_{js}(P\|Q)$ is convex in P.*

Proof. $D_{js}(P\|Q)$ is a sum of two convex functions, and therefore convex itself. □

Theorem 3.6. *If Q is fixed, then $D_{kl}(P\|Q)$ and $D_{js}(P\|Q)$ are maximized by P that is δ shaped (i.e., $P = [0,\ldots,0,1,0,\ldots,0]$).*

Proof. By Kuhn-Tucker conditions, since $D_{kl}(P\|Q)$ and $D_{js}(P\|Q)$ are convex in P, their maximum is achieved at the margins, i.e., $P = [0,\ldots,0,1,0,\ldots,0]$($\delta$ vector) [13]. □

In Theorem 3.6 we showed that the content distribution that maximize the divergence measure of a product is *delta* shaped. Let us go back to the marketing theory and explain this result. In small-world modeling, a δ spatial distribution of sales can be attributed to the following conditions: First, there is no national level advertisement, namely, no external influence; second, virtually no random links (no weak ties) between individuals in distant geographic regions; and finally, a single initial location where some sales do occur (due to local exposure). If the product is successful, the WOM effect will take place in the vicinity of the first adopters. Sales will increase only in the region where the first sales were made and all the other regions will have no sales at all. The divergence measure suggested by Garber et al. [11] will therefore, give a higher value to a distribution in which all the sales occurred in the same region, than to a distribution with the anticipated adopter clusters. This may mean that the use of the divergence in [11] is not optimal. We thus suggest looking into other pattern recognition methods to farther improve the results reported by Garber et al. [11]. This, however, is beyond the scope of our work.

Interestingly, when emerging artists are considered, a δ-distribution caused by the above conditions, is the typical case. For example, a new rapper usually starts his career by performing in her local neighborhood. If she is successful her initial audience will spread the word, and the artist will become increasingly popular in the region of her hometown. However, as long as the rapper does not have the means for a national level campaign, she is very unlikely to break out of her original region. Only after she is signed by a major record company, she has the means to gain nation wide popularity. Human scouts try to find these local artists before a competing company approaches them. Our algorithm aims at doing the same, by detection these delta shaped popularity distribution as reflected in a P2P network.

3.4. Divergence in P2P queries

In practice, sales data is not extracted from uniform sized regions. For example, US sales may be aggregated on the level of states. More sales in California than in North Dakota, may not necessarily mean that the product is received better in California since the population size difference between the two states is huge. Therefore, if the distribution of customers is not uniform, we need to adjust the divergence measure by letting the vector Q reflect the spatial distribution of potential customers at each region.

This study investigates the popularity of new artists according to their local popularity as reflected in the Gnutella network. Our spatial distribution is based on US metropolitan areas (cities), according to the geographical resolution of query strings. The cities differ in the amount of total query strings originating from them. Thus, in our divergence measurements, we set the vector Q to represent the distribution of query strings in the classified regions. Component $Q(i)$ is the fraction of query strings originated from region i out of the national total number of queries:

$$Q(i) = \frac{R(i)}{\sum_{i=1}^{N} R(i)} \tag{5}$$

where $R(i)$ is the number of query strings in region i and N is the number of regions. In the reminder of the paper we will use Q as the distribution vector of all the intercepted queries (as we explained above), and P for the distribution of queries specific to some artist or song, namely

$$P(i) = \frac{R_s(i)}{\sum_{i=1}^{N} R_s(i)} \tag{6}$$

where $R_s(i)$ is the number of queries in region i that are in the subset of strings s relating to a ceratin artists, and N is the number of regions.

We experimented with both forms of divergence measurements, and found them both similarly useful. The KL divergence have a higher range of values, making it easier to discern delta shaped distributions. On the other hand, JS divergence measurements are more steady at times when the distribution is based on a small number of queries. KL divergence tend to have more fluctuations (as a result of the wider range of values). To avoid confusion, we limit our discussion from here on to the Kullback–Leibler divergence.

4. Spatial diffusion of emerging artists in P2P networks

In this section, we investigate the diffusion of songs created by new artists in time and space. We show that emerging artists can be seen as a special case of product innovation in which the main growth factor is the word of mouth (WOM) effect, and the external factors are negligible. We examine the local popularity and the distribution of queries before a breakthrough at the national level occurs. Using marketing terminology we can say that a good "product" means a catchy hit single, and that the lack of a nation wide advertising campaign results in a δ-distribution of fans, namely almost all the fans are coming from the same geographical location. This δ distribution is reflected in high divergence values. We also show that after a commercial breakthrough, the artist reaches a larger audience and the spatial distribution of queries becomes close to uniform resulting in a decreased divergence.

The *Shop Boyz* are a typical example of locally popular artists rise to nation wide success. Emerging from the

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

Table 1
Atlanta's local popularity chart on February 18th (week 8).

Position	String	Frequency
1	Adult	882
2	Akon	583
3	Lil wayne	345
4	This is why im hot	290
5	Justin timberlake	270
6	Fergie	233
7	Beyonce	230
8	Dont matter	229
9	Mims	224
10	Pretty ricky	203
11	*Party like a rockstar*	198
12	Ciara	195
13	Porn	186
14	*Party like a rock star*	185
⋮		⋮
37	*Shop boyz*	132

Bankhead area of Atlanta, their hit single *Party Like a Rockstar* entered the *Billboard Hot 100* on the chart issued on May 5th 2007 (week 18) at the 80th position. On the chart issued on June 9th (week 23) it already reached the second position. *Party Like a Rockstar* became to be a world-wide Hip-Hop anthem and MTV's *Summer Hit of 2007*. This song is ranked 14 on the Billboard's year-end *Hot 100* chart of 2007.

4.1. The Shop Boyz example

On February 18th 2007 (week 8), very few people outside Atlanta knew who the *Shop Boyz* were. The string "party like a rockstar" ranked 10,156 on our global queries chart. However among the Hip-Hop fans in Atlanta the group was already highly popular. Table 1 shows the local popularity chart in Atlanta that week. The strings "party like a rockstar" and "party like a rock star" are ranked 11 and 14, respectively. The string "shop boyz" was ranked 37 in Atlanta that week. This is especially impressive considering the fact that, as we describe later, the song entered the bottom of the Billboard charts only a month and a half afterwards. Also note that the query chart contains the strings "adult" and "porn" which are not music related. By removing the non-music related queries, we get even higher rankings for the song.

Table 2 shows the total number of queries intercepted by our system in eleven major US cities, the number of *Shop Boyz* related queries,[3] and the corresponding distribution vectors Q and P. This information reflects the data sampled on the weekends of February of 2007. Note the absolute dominance of Atlanta in the *Shop Boyz* related queries. Over 90% of the queries originated from there. The Kullback-Liebler divergence between P and Q, is 2.4 which is relatively high. For example, the KL divergence of the popular string "adult" was only 0.02 and the divergence of the string "Avril Lavigne" (a well establish artists) was 0.0123 during that

Table 2
Number of total queries, number of Shop Boyz related queries and the corresponding distribution vectors Q and P for different US cities. The data was sampled on the weekends of February 2007.

City	All queries	Shop Boyz	Q	P
Atlanta	778,960	1046	0.123	0.906
Chicago	761,124	6	0.12	0.005
Dallas	875,189	11	0.138	0.01
Detroit	262,193	2	0.041	0.002
Houston	644,150	10	0.102	0.009
Los Angeles	446,822	55	0.07	0.048
New York	859,056	5	0.135	0.004
Philadelphia	284,607	7	0.045	0.006
Phoenix	518,578	4	0.082	0.003
San Antonio	454,705	3	0.072	0.003
San Diego	455,547	6	0.072	0.005

time. When mapping the geographical location of the *Shop Boyz* audience, the popularity in Atlanta is reflected as an almost δ shaped distribution vector. This δ distribution is an indication of a large base of local audience. A cluster of early adopters passing the word of the hottest new band in town. Since the *Shop Boyz* signed with Universal Republic only several weeks later,[4] there was very little advertisement outside Atlanta that could have affected potential audience.

Fig. 2 depicts *Party Like a Rockstar* popularity and the KL divergence measure. The popularity was measured by calculating the percentage of "*Party Like a Rockstar*" related query strings of the total number of queries sampled on that week. The divergence was calculated for every week as described above. The fluctuations in the first few weeks (January) are due to the limited number of *Shop Boyz* related queries intercepted in these weeks: a total of 132 in the first four weeks (compared with the 1,046 intercepted in the four weeks after). The two obvious trends are the increase in the band's popularity and the decrease of the divergence. The band started to gain popularity in February, and on the 11th week of 2007 the divergence started to decline significantly. Again, this is 7 weeks before it even entered the *The Billboard Hot 100*. If you lived the Hip-Hop scene in Bankhead Atlanta, you might have known the *Shop Boyz* were hot already in early 2007, but for the rest of us, this information was not available. Using our algorithm from Section 5 with the right parameters (local threshold $T_l = 500$ and the second detection pattern), one can spot out the string "party like a rockstar" in the local P2P popularity chart of Atlanta in the 6th week of 2007 (February 4th).

4.2. Additional examples

Fig. 3 depicts the popularity and divergence of four artists. In Fig. 3a we see an already known, well established artists: *Madonna*. As expected, the divergence values are low, the popularity values are high, and both graphs maintain relatively constant values. Fig. 3b and c are two more examples of emerging artists: *Yung Berg* from Los-Angeles and *Soulja Boy* from Atlanta. In both figures we see the in-

[3] *Shop Boyz* related queries are the queries that include the substrings "shop boyz" or "party like a rockstar". The matching operation was case insensitive.

[4] Universal Republic is part of Universal Music Group. The deal was announced April the 10th, 2007.

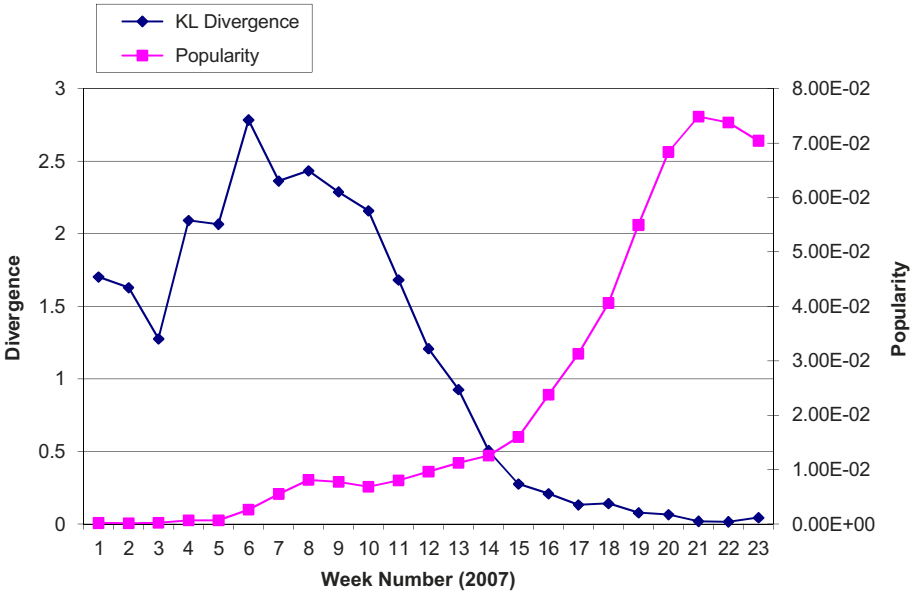


Fig. 2. *Party Like a Rockstar* popularity and KL divergence.

crease in popularity is simultaneous to the decrease in the divergence (in Fig. 3c these trends are evident after week 20). The detection algorithm suggested in the next chapter spots both of these artists many weeks before they enter the bottom of the Billboard's charts. Fig. 3d depicts an additional locally popular artists: *Mistah F.A.B.* from the Bay area. At least up to the end of our measurement period, this artist remained only locally popular, as indicated by the low popularity values and the high divergence. According to different hip-hop websites, his success was stemmed after he faced serious media criticism due to his controversial lyrics [14]. This example demonstrates that local popularity alone, is not a sufficient predictor for artists future success. We shall further discuss this in Section 5.

5. A scouting algorithm

Based on the observations in Section 4, we devised a scouting algorithm for detecting query strings belonging to emerging artists. The algorithm is designed to be executed on weekly intervals. Each week the algorithm scans the new list of geo-aware query strings, and outputs a short list of query strings with high probability to belong to emerging artists.

In the previous sections, we saw that emerging artists are characterized by high divergence values. Thus, one might conclude that the divergence measure itself is a sufficient indication for an emerging artist. In fact this would have been in accordance with the work of Garber et al., where it was shown that high values of divergence in innovations indicate higher probability to succeed [11]. In our case however, high divergence in itself is insufficient: Whereas in [11], a product's geographical distribution is used in order to predict its success probability, in our case not every query string is a "product". One example is the case of rare spelling mistakes and typos, where the distribution vector of the query string P would be a perfect δ vector. The divergence value will be maximized, but obviously this string does not represent an emerging artist. High divergence is thus a necessary but not a sufficient condition, as it only considers the geographical distribution of the queries while disregarding their local popularity strength. Moreover, computing divergence (either KL or JS), for millions of different strings requires considerable computing resources and time and thus impractical for large scale data-sets. The problem we faced is therefore, more complex than in [11], and might be conceptualized as looking for a "needle in a haystack".

5.1. Patterns of local and global popularity

Our approach is based on trying to spot artists that are extremely popular in one geographical location, while still unknown in the rest of the US. Such artists will have a δ shaped distribution of content, but with additional constraint on the minimum *local popularity*. We model *local popularity* of queries using T_g and T_l – global and local popularity thresholds. Suppose $r_g(i)$ and $r_l(i)$ are a query's global and local charts ranking at week i, respectively. In our study, $r_g(i)$, is the ranking of a query in the US queries popularity chart, and, $r_l(i)$ is the ranking in a city's popularity chart. We require queries of emerging artists to obey:

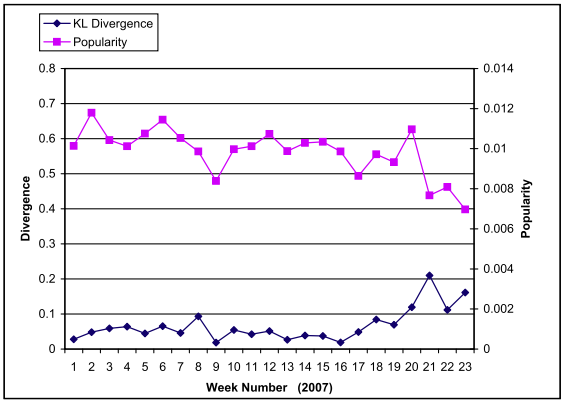
$$r_l(i) \leqslant T_l \tag{7}$$

$$r_g(i) \geqslant T_g \tag{8}$$

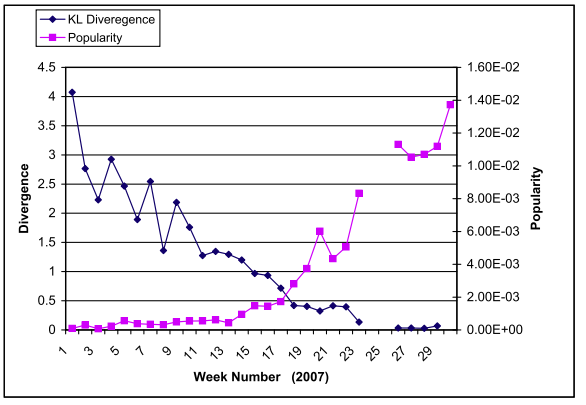
and

$$r_l(i) \leqslant r_g(i) \tag{9}$$

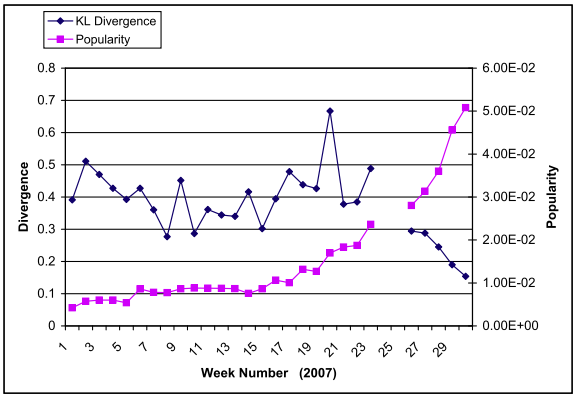
The first condition (7) assures a minimum level of local popularity, meaning the artist has a stronghold of hometown audience. The second condition (8) assures that the



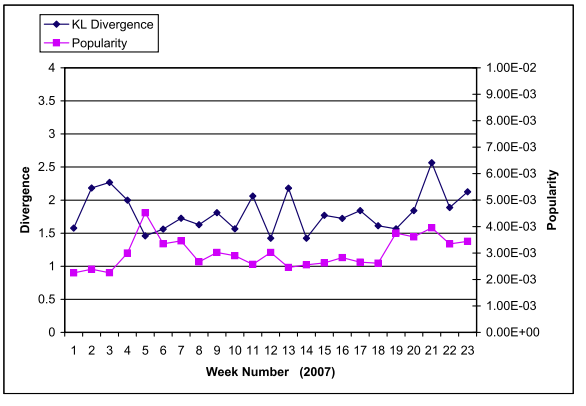
(a) Madonna



(b) Yung Berg



(c) Soulja Boy



(d) Mistah F.A.B.

Fig. 3. Popularity and KL divergence – different showcases. Samples of weeks 24 and 25 are missing due to technical difficulties, as explained in Section 2.1.

artist is not globally popular, and the third condition (9) requires that the local popularity ranking will be higher (lower in value) than the global one. It thus follows that choosing T_g and T_l such that:

$$T_g \geqslant T_l \tag{10}$$

and maintaining (7) and (8) assures that a query is *locally popular*.

The global popularity threshold, T_g, is used to distinguish the top most popular queries in the US, from the rest of the queries. Each of our weekly global popularity charts for US queries, consists of 1.73 million strings on average. Based on our observations, we wanted the global popularity threshold to approximately discern the top one thousandth of the chart. We thus chose $T_g = 2000$. An artist that succeeds in entering the global top two thousand queries list will experience thousands of downloads a day. For example, in the sample taken on January the 14th (week 3), at the bottom of the top two thousand chart are the strings "jimmy hendrix" and "ipod movies" both with 567 unique US based geographically identified queries. Obviously these strings belong to popular queries. Remembering that approximately 75% of the queries were removed after traveling more than two hops, and taking into account the other pre-processing filtering as described in Section 2, we can assume these searches were performed at least thousands of times a day. Formulating the three conditions in (7), (8), and (10) allowed us to eliminate rare queries (7), while still maintaining the demand for non-uniform distribution: (8) and (10).

5.1.1. The All Times Popular List (ATPL)

In order to detect new artists, one needs to filter out already famous artists and non-relevant query strings, e.g., sex related (22%), movies (4.1%), software (0.54%), etc. We used the data collected in 2006 in order to create the initial *All Times Popular List*, henceforth *ATPL*. This list is comprised of all the strings that reached the global top T_g queries sometime during the history, in 2006. Obviously it contained many popular strings related to pornography such as: "sex", "adult", and "porn". Unlike emerging artists, these strings are typically "non-volatile", having constant popularity over time [6]. *ATPL* also contains many artists that were already popular before 2007 e.g., "jay z", "akon", "madonna", and "avril lavigne". By ignoring the strings in

ATPL, we filter out "non-volatile" queries that typically relate to well established artists and pornography.

Naturally, as time goes by, new strings become popular, and *ATPL* needs to be updated. As described above, our algorithm is iterative, and the *ATPL* is passed from one iteration to the next. $ATPL_0$ is the initial data collected in 2006. At iteration i, when examining new queries data, the algorithm first updates $ATPL_{i-1}$, from the previous iteration by adding all the new strings that passed the T_g threshold in the current global popularity chart. The updating is done without removing any of the strings already in the list, therefore $ATPL_i$ is an aggregation over the entire history until the examined time.

5.2. Algorithm description

The algorithm's inputs for week i are all the geo-aware queries collected since the previous execution; local popularity charts from previous iterations; and $ATPL_{i-1}$. The outputs of the algorithm are a list of query strings with high probability to belong to emerging artists; local popularity charts for week i; and $ATPL_i$. The latter to be used by the algorithm in week $i+1$.

Algorithm 1. Pseudo-code using pattern 2

```
build G_CHART_i;
build L_CHART_i;
ATPL_i = ATPL_{i-1} ∪ G_CHART_i[1,...,2000];
for position ← 1 to T_l do
  if L_CHART_i[position] ∉ ATPL_i then
    // pattern 2:
    if (r(2) − r(0)) > 0 then
      output string;
    end
  end
end
```

Algorithm 1 presents a pseudo-code of the algorithm. The first step is to compile the global and local popularity charts (G_CHART_i and L_CHART_i respectively). The local popularity chart is trimmed at T_l. By doing so, we assert condition (7). For each such query string, we first filter against $ATPL_i$, and thus asserting (8). Then, the algorithm examines the local chart rank values in the past n weeks, and looks for "promising" patterns. In other words, the algorithms looks for patterns in the tuple $\overline{R_l} = \langle r_l(0), r_l(1), \ldots, r_l(n)\rangle$, where $r_l(0)$ is the local weekly popularity of the string in this week, and $r_l(j)$ is the local popularity of the string j weeks ago. If a desired pattern is found, the algorithm outputs the string. Table 3 depicts the different detection patterns tested. We discus these different patterns in Section 6.1.

6. Evaluation

We executed the iterative algorithm on the first twelve weeks of 2007. For each string that was marked by the algorithm, we checked whether it reached the global threshold T_g in one of the following weeks, until the 30th week of 2007 (when our data collection efforts ended). If the string reached *global popularity*, we classified it as a *success hit*. We thus define the *success precision* of the algorithm as the percentage of success hits of all the strings indicated by the algorithm.

Table 3
Local popularity growth detection patterns.

Pattern 1	$r_l(1) > r_l(0)$
Pattern 2	$r_l(2) > r_l(0)$
Pattern 3	$r_l(1) > r_l(0)$ and $r_l(2) > r_l(1)$
Pattern 4	$r_l(2) > r_l(0)$ and $r_l(1) - r_l(0) > r_l(2) - r_l(1)$
Pattern 5	$r_l(1) > r_l(0)$ and $r_l(2) > r_l(1)$ and $r_l(1) - r_l(0) > r_l(2) - r_l(1)$

Some artists do not follow the success patterns we target and thus cannot be discovered by our algorithm. However, no record label can handle more than a limited number of artists. Therefore, when scouting artists is considered, there is little significance to measuring the recall rate. We thus focus our analysis on optimizing the *success precision* only.

In the first twelve weeks of 2007, 1612 new strings entered *ATPL*. There are approximately 1.73 million unique strings on each week. Thus, the probability of a random pick to be a success hit is $\frac{1612}{1.73M} < \frac{1}{1000}$, which is equivalent to a precision of 0.1%. The algorithm suggested in this study has an average success percentage that ranges between 15% and 30%, which is an improvement of two orders of magnitude over a random pick and favorably comparable to the industry standard.

6.1. Local popularity growth detection patterns

Fig. 4 depicts the success percentage for the detection patterns in Table 3 using queries from five cities and $T_l = 500$. We focused on patterns not longer than three weeks ($\overline{R_l} = \langle r_l(0), r_l(1), r_l(2)\rangle$), since they are easy to implement. Results are presented for five major cities and their aggregation. All the patterns assert a recent increase in the local popularity. This means a negative local derivative in the ranking tuple $\overline{R_l}$, since higher popularity translates to a lower chart position. Also presented is the algorithm's success percentage when it uses no history at all. In this case detection is based on the current local popularity rank alone. For the five cities aggregated, this simple "no history" pattern already gives us a success percentage of 14.4%.

The first pattern requires a popularity improvement since the previous week; namely: $r_l(1) > r_l(0)$. In this example, pattern 1 seems to reduce the overall success percentage by 2.3%. However, in measurements we performed in other cities pattern 1 showed a smaller improvement, and insignificant when compared to the other patterns. It seems that in order to track popularity increase in emerging artists, we need to look at time periods longer than one week.

The second and third patterns require a two week popularity improvement. These two patterns are effective in increasing the success percentage by more than 10%. The second pattern is a more "relaxed" version of pattern 3. While pattern 3 requires a popularity increase of two

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

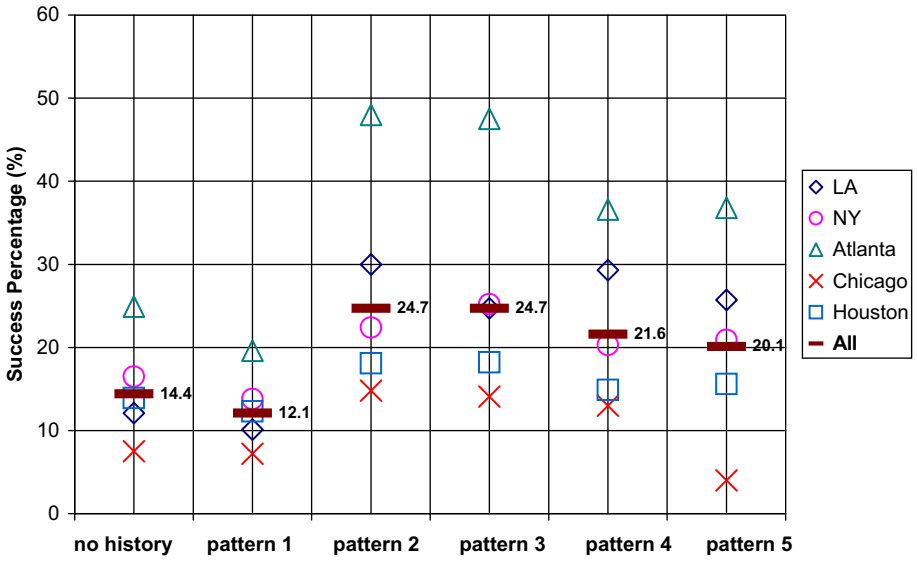


Fig. 4. Success percentage for different detection patterns.

weeks in a row, pattern 2 only requires a popularity increase since two weeks ago. When compared in other cities, pattern 2 yields slightly higher success rates than pattern 3. Again, we interpret this result by suggesting that popularity changes should be measured on time periods longer than one week and a small weekly decline in popularity should be ignored.

We attempted to improve on patterns 2 and 3 by requiring also a negative second derivative. We thus tried pattern 4 and 5 which are similar to 2 and 3 with one additional constraint: $r_l(1) - r_l(0) > r_l(2) - r_l(1)$. This means that not only the chart position is higher (lower in value) from two weeks ago, but also the rate of climbing up the chart accelerates. Fig. 4 shows that this attempt failed. Apparently, when a song is climbing up the P2P popularity chart, it usually makes bigger leaps forward when it is still in the lower part of the chart, but when it reaches the top of the chart progress seems to come in smaller steps. This means that the leaps up the chart are getting smaller, and the second derivative is actually positive. It should be understood that in most cases of promising artists, the increase in the number of new queries per week accelerates, but as ones moves up the chart the gap between successive positions grows (especially in the top positions), thus it is harder to translate this increase to an advance in ranking. This gap increase is inherent to power law distributions which was found to be common in many P2P queries distributions [4]. We test this assumption by plotting the queries distribution in our data-set, and indeed the results in Fig. 5 depict a distinct power law distribution of queries popularity. We conclude that for higher values of T_l pattern 2 shows the best performance. Therefore, we will discuss the remaining results using this pattern.

Fig. 4 shows that some cities have higher success rate than others. In the early nineties Seattle's music scene was considered very "hot", as many famous *Grunge* bands came from the area (*Nirvana*, *Pearl Jam*, *Alice in Chains* and others). Today, Atlanta is considered "hot" in the Hip-Hop scene as reflected by its high success rate. Back in the nineties, it was not an easy task for a record company based in New York, to follow the Grunge scene in Seattle without actually being present. Today this can be done from anywhere in world, simply by monitoring the P2P activity.

Fig. 6a depicts the weekly average success percentage for different local threshold values (T_l) in five cities (using pattern 2). Obviously, when $T_l = 0$, the algorithm detects no emerging artists, since it does not consider any of the strings. When $T_l < 50$, the success percentage is relatively low. Since locally popular emerging artists usually do not reach the very top of the local chart, the algorithm misses too many of them. However, in most cities when $T_l > 50$, the success percentage increases dramatically, and a maximum is reached in the range of $50 < T_l < 500$, where the peak changes from city to city. The average curve peaks when $T_l = 400$ with a success percentage of almost 30%. For $T_l > 500$, condition (7) is relaxed and the success percentage decreases. Higher values of T_l will cause the algorithm to detect the artists earlier, but with less certainty. Chicago is an exception since its success percentage increases in this range and reaches a maximum at $T_l = 1500$. We explain this by suggesting that the local artists in this city do not manage to gain enough popularity to reach higher values in the local chart, and thus languish at the lower positions.

Fig. 6b depicts the aggregated number of unique strings that the algorithm indicated during the first 12 months of 2007 (upper curves), and the aggregated number of actual success hits (lower curves). Obviously the number of indicated strings will always be higher or equal to the number of actual success hits. As the local threshold increases, the algorithm reviews more strings, and thus all curves are monotonically increasing. However, the number of indicated strings increases at super-linear rate, while the number of actual success hits increases much slower. The latter

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

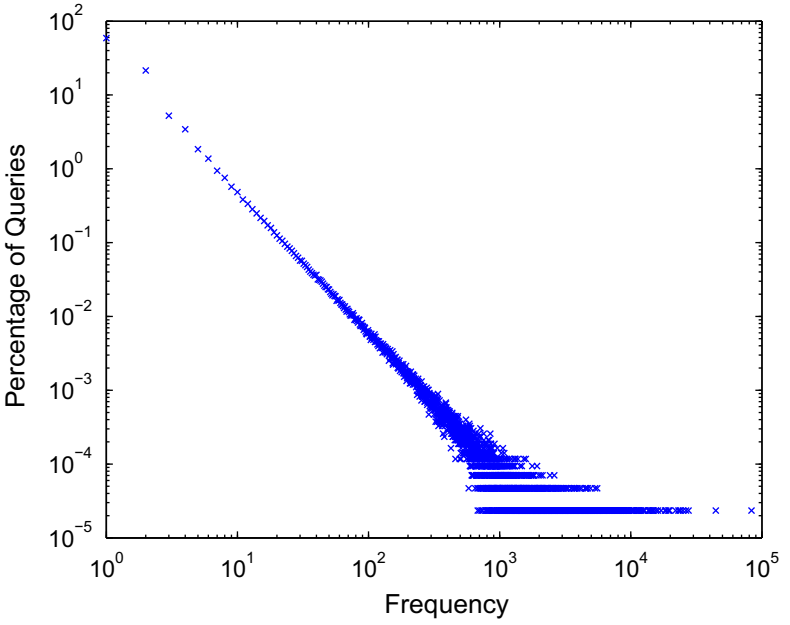
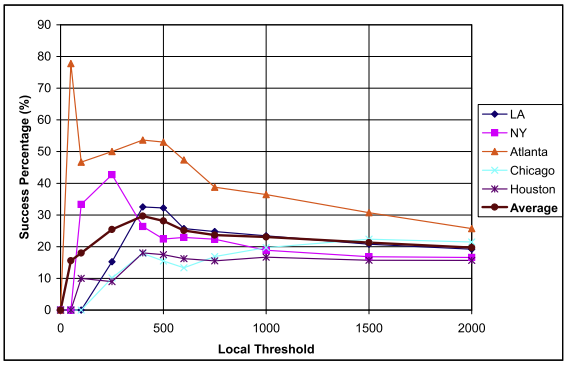
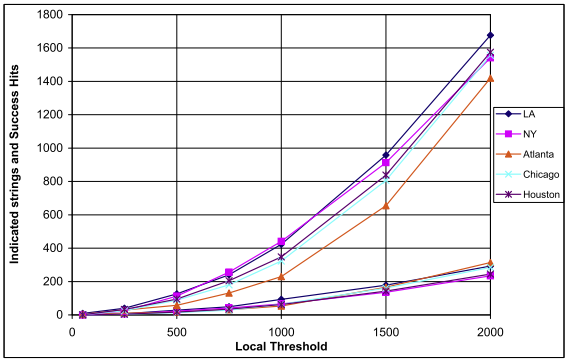


Fig. 5. Power law distribution of queries in the Gnutella network.



(a) Success Rate vs. Local Threshold



(b) Identified Strings and Success Hits vs. Local Threshold

Fig. 6. The effect of T_l on the success percentage.

curves show over 200 different success hits during that time period. However, the number of identified artists in this time period is much smaller since most artists have more than one success hit related to them. For example, when $T_l = 1000$ the *Shop Boyz* have five different success hits in Atlanta (The strings: "shop boyz", "shop boys", "party like a rockstar", "party like a rock star" and "like a rock star"). Finding the optimal local threshold for each city, as well as the best detection pattern can be further fine-tuned.

6.2. Detection time

The time gap between the algorithm's detection time and an artist's actual success is of special interest for the record laels. An algorithm that can predict global popularity only a week or two before it actually occurs, can be useful, but limited in its commercial value. The artist is probably already signed by a national-level recording company. However if the artist is spotted a month or two before its breakthrough, it usually gives a recording company plenty of time to negotiate a deal. Therefore, for each string that was categorized as a *success hit*, we measured the time gap between the week it was first detected by the algorithm, and the week it first reached global popularity. Results are presented in Fig. 7 for detection pattern 2. The graph aggregates results from 11 major cities in the US (the same cities as in Table 2). The blue line indicates the average time gap in weeks for different local thresholds, while the bars indicate the number of *success hits* considered. From Fig. 7 it is apparent that the time gap varies

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021

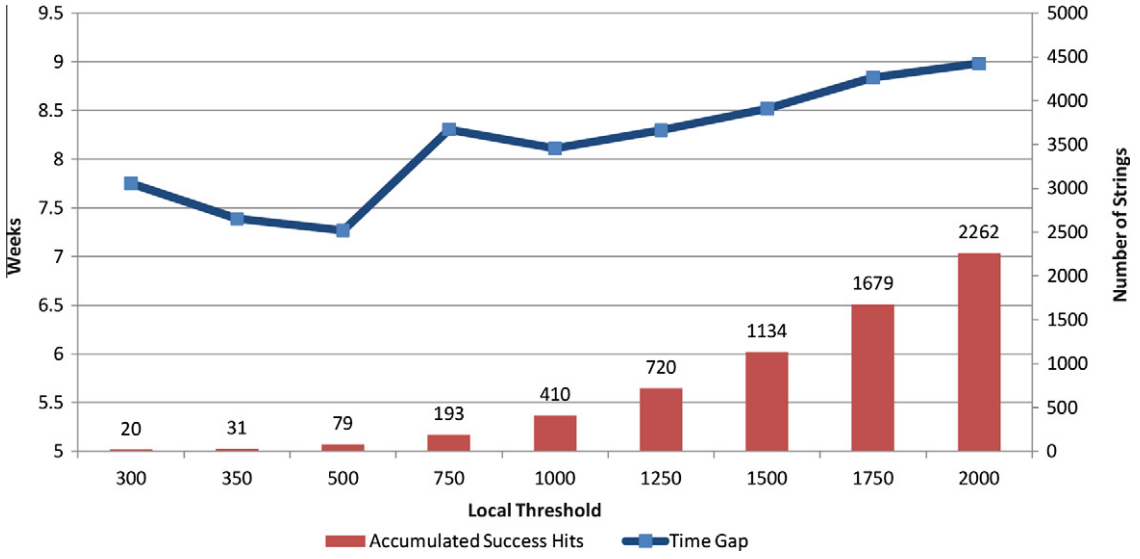


Fig. 7. Time gap between detection and actual success.

between 7 and 9 weeks. When local threshold is low, the number of *success hits* considered is limited, and thus the time gap is a little flickery. However, as the local threshold increases, the time gap graph stabilizes, and a trend of slow linear increase appears. This increase is explained by the fact that when higher local threshold is used, the algorithm considers strings from lower positions in the local chart, and thus detecting the artists earlier.

6.3. Human inspection

We cross-validated the results above by manually investigating the output of the algorithm. Manual cross-validation is a highly laborious task. It requires resolving the meaning of many locally popular search strings that ranked many months ago. We therefore focused our efforts on Atlanta, the city that scored best in all the automatic verifications above. Every week the algorithm had 8.16 predictions on average. Over the entire 12 weeks period, the algorithm had 98 predictions, 58 of them unique. We were able to manually resolve the meaning of 50 of these strings (86.2%). Fig. 8 depicts these results. Of which, 23 strings relate to 14 different musical artists that were locally popular in Atlanta during that time period. Artists that had their debut single in the Billboard only after the algorithm detection week, were marked as *successful prediction* (36%). Artists that had a debut single in the Billboard before the prediction time, were marked as *well established* artists (36%), and artists that had no singles in the Billboard up to the day of writing, were marked as *local artists* (28%). For these predictions, we show in Table 4 the detection time and the week they entered the Billboard. The average time difference is 14.2; a longer time difference than we had found in Section 6.2. However, we already showed in [7,15] that songs tend to become popular in P2P networks before, they reach their top rank in the Billboard, and that songs' popularity trends on the Billboard follow popularity trends in a P2P network, which explains the longer time difference.

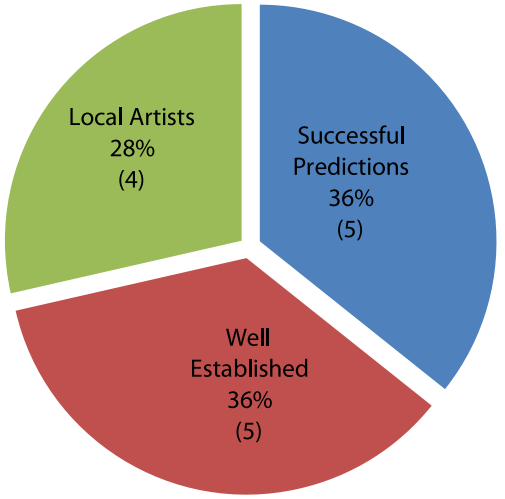


Fig. 8. Manual classification of artists' names from Atlanta. The artists names were taken from the query strings that relate to musical content.

Table 4
Successful predictions in Atlanta 2007.

Artist	Single	Detection week	Debut week	Δ
Hurricane Chris	*A Bay Bay*	9	25	16
Gorilla Zoe	*Hood Nigga*	7	15	8
Shop Boyz	*Party Like a Rockstar*	6	18	12
Baby D	*I'm Bout Money*	6	11	5
Soulja Boy	*Crank That*	Late 2006	30	30+

7. Conclusions

We model the popularity diffusion of emerging artists in time and space, and show how promising artists are characterized by high divergence values, that indicate a δ shaped distribution of fans. As the artist gain nation wide popularity her audience distribution approaches a uniform distribution which lead to near zero divergence values. Based on these observations we mathematically model *local popularity*, and design an algorithm to detect locally popular artists in a database of geo-aware P2P queries from the Gnutella file sharing network. The algorithm quantifies unsigned artists' popularity anywhere in the world, and predicts success with precision rates that are well above the industry standards. Some of our ideas are already utilized by record companies to identify unsigned artists, and our behavioral model and spatial divergence based techniques can be applied to other social based local phenomena detection.

References

[1] D. Weissman, The Music Business: Career Opportunities and Self-Defense, Three Rivers Press, New York, 2003.
[2] P. Resnikoff, Digital media desktop report: fourth quarter of 2007, 2008. Digital Music Research Group.
[3] A. Klemm, C. Lindemann, M. Vernon, O.P. Waldhorst, Characterizing the query behavior in peer-to-peer file sharing systems, in: Internet Measurement Conference, Taormina, Italy, 2004.
[4] K. Sripanidkulchai, The popularity of gnutella queries and its implications on scalability, 2001. Featured on O'Reilly's website <www.openp2p.com>.
[5] S. Meng, Y. Shao, C. Shi, D. Han, Y. Yu, Mining and predicting duplication over peer-to-peer query streams, in: ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining, IEEE Computer Society, 2006, pp. 648–652.
[6] A.S. Gish, Y. Shavitt, T. Tankel, Geographical statistics and characteristics of p2p query strings, in: The 6th International Workshop on Peer-to-Peer Systems (IPTPS'07), 2007.
[7] N. Koenigstein, Y. Shavitt, N. Zilberman, Predicting billboard success using data-mining in p2p networks, in: Proceedings of the 2009 11th IEEE International Symposium on Multimedia, 2009.
[8] M.S. Granovetter, The strength of weak ties, The American Journal of Sociology 78 (1973) 1360–1380.
[9] J.J. Brown, P.H. Reingen, Social ties and word-of-mouth referral behavior, The Journal of Consumer Research 14 (1987) 350–362.
[10] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440–442.
[11] T. Garber, J. Goldenberg, B. Libai, E. Muller, From density to destiny: using spatial dimension of sales data for early prediction of new product success, Marketing Science 23 (2004) 419–428.
[12] L.A.N. Amaral, A. Scala, M. Barthelemy, H.E. Stanley, Classes of small-world networks, PNAS 97 (2000) 11149–11152.
[13] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, 1991.
[14] J. Shepherd, Ghost rider fallout haunts mistah F.A.B., Featured on The VIBE Magazine (2007).
[15] N. Koenigstein, Y. Shavitt, Song ranking based on piracy in peer-to-peer networks, in: Proc. International Symposium on Music Information Retrieval, 2009.

Noam Koenigstein received his B.Sc. in Computer Science (cum laude) from the Technion – Israel Institute of Technology, Haifa, Israel in 2007, and his M.Sc. in Electrical Engineering from Tel-Aviv University, in 2009. Currently he is a Ph.D. candidate in the School of Electrical Engineering at Tel-Aviv University. His research is focused on Large Scale Multimedia Information Retrieval and Recommender Systems.

Yuval Shavitt received the B.Sc. in Computer Engineering (cum laude), M.Sc. in Electrical Engineering and D.Sc. from the Technion – Israel Institute of Technology, Haifa, Israel in 1986, 1992, and 1996, respectively. After graduation he spent a year as a Postdoctoral Fellow at the Department of Computer Science at Johns Hopkins University, Baltimore, MD. Between 1997 and 2001 he was a Member of Technical Staff at Bell Labs, Lucent Technologies, Holmdel, NJ. Starting October 2000, he is a Faculty Member in the School of Electrical Engineering at Tel-Aviv University, Israel. His research interests include Internet measurements, mapping, and characterization; and data mining peer-to-peer networks.

Please cite this article in press as: N. Koenigstein, Y. Shavitt, Talent scouting in P2P networks, Comput. Netw. (2011), doi:10.1016/j.comnet.2011.10.021