Category dynamics and cluster spanning during the emergence of the Lebanese newspaper industry (1851–1879)

Najib A. Mozahem

College of Business Administration, Rafik Hariri University, P.O. Box: 10 Damour – Chouf 2010, Lebanon

* Corresponding author.
E-mail addresses: mozahemna@rhu.edu.lb, najib.mozahem@gmail.com (N.A. Mozahem).

Abstract

Recently, researchers have started to pay more attention to a usually ignored topic: audience perceptions. Legitimacy, for example, is no longer modeled as the number of organizations in a population. It is now thought to be dependent on how audience members perceive these organizations. This paper will study how the newspaper industry in Lebanon emerged. The paper studies the period 1851–1879, building on the theoretic formulation of Hannan et al. (2007). The concept of cluster formation will also be introduced in order to help answer the question of whether unified identity projection is a necessary condition for successful legitimation and emergence. So far, research has produced diverging results as to the necessary conditions for successful legitimation. Cluster Analysis is used to show that in the case of the Lebanese newspaper industry, successful emergence was attained without the need to project a unified identity. In fact, the analysis clearly shows that there were two separate groups of clusters that had emerged by the end of the period. The nature of these two clusters will be investigated by looking at the category spanning activities of the newspapers that were members of the clusters.

Keywords: TBC, Business, Industry, Information science
1. Introduction

Research has shown that identities play a central role in population dynamics (Carroll and Swaminathan, 2000; Liu and Wezel, 2013; Pozner and Rao, 2006), in product success (Jensen, 2010; Khessina and Carroll, 2008), and ultimately in the success of the firm as a whole (Carroll et al., 1996). Previous identity formulations (Albert and Whetten, 1985) viewed identity as being determined solely by internal members (Hsu and Hannan, 2005) or by internal enduring attributes (Wry et al., 2011) and features (Hannan et al., 2006). Recent theoretical advances made by Pólos and colleagues (Hannan et al., 2007; Pólos and Hannan, 2001, 2002) have argued that the identity of the organization consists of the social codes that audience members use for the categories of which it is a member. Using this new theoretical foundation, organizational ecologists have shifted their attention to the period of industry formation. With this shift, increasing attention has been directed toward the unresolved topic of legitimacy (Wry et al., 2011). According to Berger and Luckmann (1967), “Legitimation ‘explains’ the institutional order by ascribing cognitive validity to its objectivated meanings. Legitimation justifies the institutional order by giving a normative dignity to its practical imperatives” (p. 93). Previously, it has been assumed that legitimacy is dependent on the number of organizations in the industry (Carroll and Hannan, 2004). However, some research has shown that small nascent populations were able to gain sufficient legitimacy (Wry et al., 2011). The new theoretical formulation of Pólos and colleagues argued that while the number of organizations in an industry mattered, a new variable, termed the “contrast”, was also crucial. If a category had high contrast, then the category would be legitimated, hence becoming a form (Hannan, 2010). Therefore, according to Hannan et al. (2007), organizations in an emerging industry need to project a unified identity.

Recent research has been able to show that identity plays a central role in the emergence and legitimation of organizational forms (McKendrick and Carroll, 2001; McKendrick et al., 2003; Perretti et al., 2008). However, this research has produced diverging results as to the necessary conditions for successful legitimation. McKendrick and colleagues have argued that a unified identity is a facilitating condition for successful form emergence, while Perretti et al. (2008) have argued that the effect of identity is context dependent. They conclude that “…different domains offer alternative default settings, and the way a new organizational form emerges is in the context of interpretation and social classification of the identity of a domain as well as the identity of candidates” (p. 543). This position gained credence when King et al. (2011) showed that a unified identity was not a necessary condition for the emergence of new forms. This paper seeks to contribute to the debate by studying whether the founding organizations in the Lebanese newspaper industry projected a single or multiple identities.
2. Background

The new theoretic formulation presents two significant departures from previous theories. First, legitimation is no longer dependent on the number of organizations in a specific category. Instead, legitimation is achieved when members of a category abide by codes that the audience uses to make sense of the actions of organizations. Categories may therefore have a high number of members that nonetheless the audience has difficulty making sense of, therefore causing the category to have what is called “low contrast”. Categories in which members follow similar rules manage to distinguish themselves from their surroundings and therefore achieve “high contrast” (Hannan, 2010). Second, unlike previous formulations (e.g., Albert and Whetten (1985)), identity is no longer perceived as being determined solely by internal members (Hsu and Hannan, 2005) or by internal enduring attributes (Wry et al., 2011) and features (Hannan et al., 2006). Ultimately, the identity of the organization will consist of the social codes audience members use for the categories of which the organization is a member (Negro et al., 2010; Pólos et al., 2002). Previous studies have shown that identity plays an important role in many dynamics. Recent research has shown that identity, not quality, is what drives resource partitioning in established markets (Carroll and Swaminathan, 2000; Liu and Wezel, 2013) and that the identity of the organization plays a pivotal role in the success of the product (Jensen, 2010; Khessina and Carroll, 2008) and inevitably the firm as a whole (Carroll et al., 1996). More importantly, research has shown that the emergence of new organizational forms, and hence their legitimation, is heavily dependent on the identity of the entrants, not on their total count (McKendrick and Carroll, 2001; McKendrick et al., 2003).

Producers, especially in nascent markets (Santos and Eisenhardt, 2009), attempt to redefine categories such that their product is considered the prototype (Navis and Glynn, 2010), to identify their rivals (Kocak et al., 2009), and to keep the status quo in their favor (Lounsbury and Rao, 2004). Meanwhile, consumers, especially gatekeepers (Hannan et al., 2007) and activists (Kocak et al., 2009), tend to classify products into different categories because this classification is the basis upon which audience members make choices (Beck et al., 2009; Biggart and Beamish, 2003; Khaire and Wadhwani, 2010; Lounsbury and Rao, 2004; Wei, 2005). Categories are especially important when audience members do not have full information regarding certain producers. By typecasting, audience members are able to apply defaults to these producers, thereby filling in the missing information (Hsu et al., 2009b).

Work by McKendrick and Carroll (2001) and McKendrick et al. (2003) has highlighted that the period of emergence of industries is a fertile ground for research. Since the identity of organizations is formed, in part, from categories of which they are a member (Hsu and Hannan, 2005; King et al., 2011; Wei, 2005), this means
that organizations in emerging categories need to display a unified identity (Hsu et al., 2009b; Lamont and Molnár, 2002). Failure to do so would confuse the audience (Hannan et al., 2007). With time, as audience engagement increases, organizations will have to differentiate themselves from their competitors, so there will be an increase in the subtypes of the new category (Kocak et al., 2009).

3. Hypothesis

In the work of McKendrick and his colleagues, identity is measured on a single dimension, the origin of the company. However, Perretti et al. (2008) found that the effect of identity is context dependent, so they argued that a more comprehensive approach is required. Their research led them to conclude that “…different domains offer alternative default settings, and the way a new organizational form emerges is in the context of interpretation and social classification of the identity of a domain as well as the identity of candidates” (p. 543). Unsurprisingly, therefore, recent research has argued that a unified identity is not a necessary condition for the emergence of new categories (King et al., 2011). In fact, Weber et al. (2008) have criticized the fact that initial variation in organizational forms has been under-examined. A discussion of these questions must be preceded by an elaboration of how organizational identities are determined. Smith (2011) used a conceptualization of identities where he categorized organizations into conformists and non-conformists by measuring the distance of the organization from the “mean”. Powell and Sandholtz (2011) highlighted firm-level attributes that they believed reflect distinct domains of origin and function. And King et al. (2011) determined identity by identifying elements related to theme, resources and services, and target population. Other studies looked at the “stories” told by organizations in order to define their identity (Wry et al., 2011).

This paper will study the identity of the Lebanese newspapers during the period 1851–1879 by building on the theoretical formulation of Hannan et al. (2007). As discussed in the Literature Review section, clusters will be used to study both the similarities and differences of the founding organizations. As mentioned above, current theoretical formulations claim that the identity of organizations is partially determined by the categories they span. Therefore, by studying category spanning and cluster formation, we will be looking at another dimension of identity.

Ruef and Patterson (2009) showed that category heuristics were less developed during the initial industry emergence period. Aldrich and Ruef (2006) stated that the conditions faced by the founding fathers were less secure than the conditions faced by their predecessors. Because cluster formation depends on the category spanning dynamics, I believe that during the emergence period, the early newspapers created more clusters than the later newspapers due to their haphazard category spanning
activities. Severe shifts in category dynamics will result in significant changes in the category space. As time goes by, newspapers should develop more-systematic dynamics due an increase in their knowledge of the market and due to an increase in scrutiny by external audience members. Therefore, I believe that the number of best-fit clusters during the later period is less than the number during the early period. In addition to forming a larger number of clusters, I believe that the early newspapers were more likely to span clusters than later newspapers. With time, as individual newspapers developed a better understanding of what type of content they wanted to include, successive newspaper issues tended to end up in the same cluster. This can be thought of as a reflection of expansion in knowledge and maturity on the part of newspapers.

**Hypothesis 1:** Earlier newspapers were engaged in more cluster spanning dynamics than later ones during the emergence period.

**Hypothesis 2:** With time, newspapers started forming fewer clusters during the emergence period.

King et al. (2011) have shown that organizations are best grouped into more than one form, while McKendrick et al. (2003) argued that a unified identity is an enabling factor for successful emergence. While previous research has used the organizations’ background as a determinant of identity or the attributes the organizations claim to possess, I will use category dynamics in order to study how the partial identities have formed during emergence.

**Hypothesis 3A:** The founding newspapers formed a single cluster.

**Hypothesis 3B:** The founding newspapers formed more than a single cluster.

I believe that the first newspapers spanned categories in a more haphazard manner than later ones. The fact that these early newspapers had no precedent to build upon and that category heuristics were yet to become well developed lends support to this belief. Although I expect that newspapers told the same stories and used a unified framing process, this does not mean that they were in agreement with regards to category spanning dynamics. While the problems and goals were uniformly identified, I believe that the content of the newspapers was not a matter of agreement. In fact, I do not even believe that issues belonging to the same newspaper spanned categories uniformly, because I do not believe that the individual newspapers had a clear formulation of the type of content they intended to include. With time, I believe that the newspapers were able to concentrate their efforts on certain categories that they believed reflected their identity and helped them reach their goals.
Hypothesis 4: During the emergence period, newspapers tended to enter new categories and exit old ones in a more random way than later newspapers.

4. Methodology

4.1. Grade of membership

Following the lead of Mozahem (2017), I determine the grade-of-membership $\mu$ of an issue in each category as the proportion of that issue that is dedicated to the category. Thus, for example, if a newspaper consisted of eight pages in which two were dedicated to sports, then the grade of membership of that issue in the category sports would be $2/8$ or $0.25$. The main advantage of this method is that this measure takes into account the engagement of the producer in each category. The greater the space dedicated to a certain topic, the higher the grade-of-membership. The reason I use the space dedicated to each category as a surrogate for the identity is that space in newspapers is a zero sum game. The more space is dedicated to one category, the less space can be dedicated to other categories. This means that when a newspaper chooses to assign a large space (which is the primary resource for a newspaper) to a certain topic, then this topic is considered to be more important to that newspaper than other topics. This was especially true during the period of emergence, because most newspapers were made up of four pages at the time because the founder of the newspaper happened to be the author of most of the articles that appeared in it. Since this resource, the space, was very limited, the choice of what to include and what not to include is a very strong indicator of the identity of the newspaper. This raises the point of whether space is a reflection of the identity as projected by the newspaper itself or as seen by the audience members. While the two are not the same, I believe that in the case of newspapers during the emergence period, they do coincide. The reason is that the product is a new physical product. When audience members are exposed to something that is physical and new to them, then the primary source of information is the attributes of that product, since, “In the simplest case, audience members recognize similarities among a cluster of producers and come to regard these similar entities as members, to varying degrees, of a common (fuzzy) set and agree on a label for the set” (Hsu et al., 2011). In addition, Hannan (2010) has stated that some previous studies have used the strategy of making “inferences about likely perceptions of audience members from other observable features of producers/products.”

It is important here to justify the choice of categories. As noted by Mozahem (2017), the justification is more practical than theoretical. As I was going through the newspapers, I noted the nature of the news stories and started classifying them. Some categories, such as economics and politics, were expected in advance, but other items came as a surprise. For example, I was surprised to see that many newspapers...
dedicated a significant portion of their space to literary issues related to the Arabic language. This led me to decide to create a separate category for such items because their frequency meant that it would be unwise to disregard them by adding them to another, more general category. During the process, I was able to identify nine different categories: Politics, Economics, Social Issues (education, immigration, etc.), Knowledge (math, physics, etc.), Literature (Arabic grammar, poetry, novels, etc.), Sport, Art, Advertisements, and Other (any item that could not be classified into any of the above groups was added to this group).

One criticism of the above might be that what I have identified as “categories” were actually features. Hannan et al. (2007) defined “a type as a coupling of a label and a schemata that articulates a view about what pattern of feature values determine the applicability of the label” (p. 60). This would mean that if what I call categories are actually features, then what I call clusters would turn out to be the true categories. This question was addressed by Mozahem (2017), where it was shown that the correlation between these categories was actually too low for them to be features.

Next, I follow the lead of Hannan et al. (2007) and Hsu et al. (2009a) and define the niche width using Simpson’s index of dissimilarity as

$$1 - \sum_{i=1}^{J} \mu_i^2$$

This measure takes values between zero and one, with a zero indicating that the entire issue is dedicated to a single category. Increasing values indicate an increase in the diversity of the newspaper and, hence, its niche width.

4.2. Clusters

Kovacs and Hannan (2009) argued that it is not enough to study the categories spanned by organizations but that it was also necessary to take into account the structure of the category space. To do this, they came up with the concept of distances between categories. Mozahem (2017) built on their work but proposed that instead of measuring the distance between categories, it would be better if one “incorporates the category space on a second level which I refer to as clusters.” These clusters are formed from category spanning activities. They are “a representation of the category-space which helps us understand how organizations span, and dedicate their resources to, different categories.”

This paper uses the same model of clusters as that used by Mozahem (2017). Cluster analysis is concerned with studying whether observations in a data set can be summarized in terms of groups or clusters. I use Agglomerative Hierarchical Cluster Analysis since my aim is to study how clusters form from the assembly of categories, i.e. the study takes a bottom-up approach. Hierarchical Cluster Analysis has been
used in many disciplines, including astronomy, psychiatry, weather classification, and market research (Everitt et al., 2011). More importantly for this research, this technique has been used in the study of the emergence of new organizational forms in biotechnology (Powell and Sandholtz, 2011) and in health care (Ruef, 2000). According to Everitt et al. (2011), “In a hierarchical classification the data are not partitioned into a particular number of classes or clusters at a single step. Instead the classification consists of a series of partitions, which may run from a single cluster containing all individuals, to \( n \) clusters each containing a single individual” (p. 71).

Fig. 1 below illustrates the formation of clusters from categories. The figure shows that there are seven categories in total and that three clusters form from the dynamics of the organizations’ category spanning dynamics. Cluster 1 contains organizations that span categories one and four, cluster 2 contains generalist organizations that span the 6 categories, and cluster 3 contains specialist organizations that span category seven only. It is important to note here that when we say an organization belongs to cluster 1, for example, it does not mean that all organizations that span category 1 belong to cluster 1. As can be seen from the figure, generalists that span all categories, including category 1, belong to the second cluster. Also important is that cluster formation depends on the extent of the category spanning activity.

As a measure of proximity, I used the Pearson correlation measure. I used the weighted-average linkage as the linkage method. Regarding the proximity measure, Everitt et al. (2011) note that the nature of the data should determine the choice of this measure. In this data set, I am using cluster analysis on variables that refer to proportions. The variables include the proportion of the issue that is dedicated to politics, the economy, sports, and so on. These variables are all measured on the same...
scale. They lie between zero and one. In such a case, it makes more sense to talk about the correlation between the data than to talk about the distance. The variables provide an indication of the “relative profile” of the observation. With regard to the choice of the linkage method, the weighted-average method “gives equal weights to each cluster regardless of how many observations it contains. Such methods consequently tend to work better for detecting clusters of unequal size” (Hamilton, 2009, p. 358).

The next step after determining the proximity measure and the linkage method is to determine the best-fitting total number of clusters. Milligan and Cooper (1985) examined 30 rules with the conclusion that while there was no single best rule for all situations, two seemed to work most of the time. These are the Calinski and Harabasz pseudo-F index and the Duda-Hart index. This paper uses both rules.

4.3. Research design

This paper uses a variety of tools to test the hypotheses. Fig. 2 illustrates how the various models are inter-linked. Content analysis will be used to collect information from the primary data sources, which are the newspaper issues. The collected information will be used to determine how the newspaper issues spanned categories. These data will then be used in cluster analysis in order to determine the optimal
number of clusters. They will also be used in the multi-Level logistic regression in order to study whether the category-spanning dynamics of the newspapers differed significantly over time. In addition, certain variables in the multi-level logistic regression will be based on data collected from the “World of Others” (Krippendorff, 2004), which, in this case, are the bibliographical sources that contained data such as the founding date of the publications.

5. Methods

All the data collected were obtained from the Jafet library at the American University of Beirut. Unfortunately, newspaper collections are not a random sample (Riffe et al., 1998). However, this does not mean that we cannot, or should not, study such collections (Riffe et al., 1998). Given the historical nature of the data and the fact that they are collected from primary and not secondary sources, it is acceptable, and expected, that the entire collections could not be located. The size of the collection at the AUB is significant: the entire collection of 14 of the 26 newspapers is available, with a considerable proportion of another four newspapers also available.

The next step was to use content analysis to study the category spanning dynamics. This meant that I first had to identify the categories. I was able to identify nine different categories: Politics, Economics, Social Issues (education, immigration, etc.), Knowledge (science, math, etc.), Literature (poetry, novels, etc.), Sport, Art, Advertisement, and Other (any item that could not be classified into any of the above groups was added to this group).

I sampled one issue from each month for each newspaper title and went through it, determining the space dedicated to each of the above nine categories. I read through the entire issue and used a ruler to determine the area of the page, since different publications had different page sizes. I multiplied the area of the page by the number of pages to obtain the total area of the publication. I then measured the area of the news items that belonged to each of the above categories. I then divided the area occupied by each category by the total area of the publication to obtain the ratio of the publication dedicated to each category. Photos were calculated as part of the area dedicated to the news item to which they belonged. The final dataset consisted of 659 entries, with each entry representing one issue.

1. The cut-off date for the study is 1879 because starting in 1880, many Lebanese newspapers migrated to Egypt, where the political context was more favorable and allowed for more freedom. The restrictions in Lebanon were mainly due to the ascension of Sultan Abdul Hamid. This is why the total number of newspapers active in the country at that time started to decrease.
6. Analysis

6.1. Cluster analysis

I used Stata version 12 to conduct the following analysis. Initially, I pooled all the issues in the period 1851–1879 together and performed a cluster analysis using the Pearson correlation as a measure of proximity and using the weighted-average as the linkage method.

Mozahem (2015) argued that the newspaper population in Lebanon matured in the middle of the 1870s. Therefore, it would be interesting to see whether the contents of the newspapers also took on a different trajectory during that same period. To look into this issue more deeply, I divided the above analysis into two different periods. The result is Fig. 3, which shows two scatter plots.

Both plots refer to Simpson’s index of the issues, a concept that was introduced earlier in this paper. To produce the figure, I grouped the issues published in the same month together and calculated the average Simpson index and its standard deviation for each month. Both of these are plotted on the figure. I also drew a vertical line at January 1875 because it appears to be the time when we can see a difference in the graphs. To the left of the line, we see that there is significant variability in both the average and the standard deviation of the Simpson index. With time, this variability decreases. To the right of the line, we see that the monthly variability is visibly much lower than that to the left of the line.

To verify the above observation, I produced Fig. 4. This figure shows the magnitude of the difference between the consecutive monthly standard deviations. The vertical line marks the beginning of 1875, while the horizontal line helps us see that starting

![Fig. 3. Simpson’s index for newspapers between 1860 and 1879.](image)
in 1875, almost all of the points are less than 0.1, while before 1875, we see that there were much higher differences in consecutive months. These figures suggest that the year 1875 would make a good choice for a cut-off period. Therefore, I split the data into two time periods, one spanning the years 1851–1874 and consisting of 362 issues and one spanning the period 1875–1879 and consisting of 297 issues. Figs. 5 and 6 present the dendrograms for the analysis.

A look at the above figures shows that the clustering effect is much stronger in the latter period. There are 14 groups in the first period at values greater than 0.6, while there are only 5 groups in the second figure. More importantly, though, in Fig. 6, we...
can see that the smallest three groups have only 13 issues in total. This means that almost all of the newspaper issues have become part of one of the two largest groups. In addition, we can see that groups 1 and 4 are almost equal in size. What these numbers tell us is the following: in the first part of the emergence period, the

**Table 1.** Distribution of newspaper titles among the groups 1851–1874.

| Group       | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|-------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|
| Amal alJamiya | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Basheer      | 0 | 0 | 0 | 3 | 0 | 0 | 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hadiqat al-Akhbar | 2 | 0 | 0 | 0 | 1 | 13 | 71 | 2 | 8 | 1 | 18 | 0 | 1 | 1 |
| Jinan        | 39 | 3 | 0 | 0 | 1 | 0 | 15 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| MajmooFawaid | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nafeer       | 0 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Nahla        | 5 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Najah        | 1 | 0 | 0 | 0 | 0 | 2 | 19 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Nashra       | 35 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sharaka      | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| Tabezib      | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Tagadom      | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Zahra        | 1 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |

| 108 | 8 | 4 | 13 | 2 | 15 | 167 | 2 | 10 | 6 | 18 | 6 | 1 | 2 | 362 |

The numbers in bold in the very last column are the sum of each row. So they are the sum of the number of issues of each newspaper title. The numbers in bold in the very last row are the sums of each column. So they are the sum of the number of issues in each cluster group.
newspapers were similar enough to be clustered together at high correlation values, but they formed several distinct groups, each with a considerable number of members.

The next question of interest is whether the issues of the same newspaper ended up in the same group. In other words, was the content of each newspaper consistent from issue to issue? Tables 1 and 2 below show the distribution of the issues of each title among the different groups for both time periods. We can see in Table 1 that there was a tendency for the issues of the same newspapers to end up in the same group, but there were exceptions to this tendency. *Hadiqat al-Akhbar*, for example, is spread over 10. *Al-Nashra al-Osboo’iya*, on the other hand, had almost all of its issues end up in the same group. In Table 2, we see that most of the issues seemed to be homogeneous except for *al-Jinan*.

It would be useful, though, if we could summarize the tables with a single number, since this would enable us to compare the two time periods. Again, we turn to Simpson’s index, but this time, instead of studying category spanning, we look into cluster spanning. So, for example, if a newspaper had a total of 50 issues in our dataset and these issues were divided into two groups by the cluster analysis, with thirty in one group and twenty in the other. The grade of membership of the newspaper in the first group was 30/50 and that of the second group was 20/50. Again, I used the formula:

\[ 1 - \frac{1}{C_0} \prod_{i=1}^{n} \mu_i^2 \]

I calculated these values, and the results are presented in Tables 3 and 4.

**Table 2.** Distribution of newspaper titles among the groups 1875 — 1879.

| Group | 1 | 2 | 3 | 4 | 5 |
|-------|---|---|---|---|---|
| *Basheer* | 2 | 0 | 7 | 38 | 0 | 47 |
| *Janna* | 0 | 0 | 0 | 2 | 0 | 2 |
| *Jinan* | 43 | 1 | 0 | 16 | 0 | 60 |
| *Lisan* | 0 | 0 | 1 | 25 | 1 | 27 |
| *Mishkat* | 4 | 0 | 0 | 0 | 0 | 4 |
| *Moktataf* | 43 | 0 | 0 | 0 | 0 | 43 |
| *Nashra* | 31 | 0 | 0 | 0 | 0 | 31 |
| *Tabeeb* | 24 | 0 | 0 | 0 | 0 | 24 |
| *Tagadom* | 0 | 0 | 0 | 3 | 0 | 3 |
| *Thamarat* | 0 | 0 | 3 | 53 | 0 | 56 |

| | 147 | 1 | 11 | 137 | 1 | 297 |

The numbers in bold in the very last column are the sum of each row. So they are the sum of the number of issues of each newspaper title. The numbers in bold in the very last row are the sums of each column. So they are the sum of the number of issues in each cluster group.
The calculated numbers during the period prior to 1875 clearly vary from one newspaper to another, with the smallest being 0 and the largest being 0.5970985. The average value indicates that, in general, the newspapers belonged to few groups. How do we check the statistical significance of this result? To do this, I ran a simulation experiment in which I pooled the newspaper issues prior to 1875 together and randomly assigned them to 14 groups. The only condition I introduced was to keep the group sizes equal to the sizes of the groups obtained via the cluster analysis. I then calculated the Simpson index for each newspaper and the overall average. I

Table 3. Simpson index for the newspaper titles before 1875.

| Newspaper        | Simpson index |
|------------------|---------------|
| Amal alJamiya    | 0.5785124     |
| Basheer          | 0.1442308     |
| Hadiqat          | 0.5970985     |
| Jinan            | 0.5383333     |
| MajmooFawaid     | 0.4444444     |
| Nafeer           | 0.4489796     |
| Nabla            | 0.53125       |
| Najah            | 0.3576389     |
| Nashra           | 0.0540124     |
| Sharaka          | 0.5           |
| Tabeeb           | 0             |
| Taqadom          | 0.375         |
| Zahra            | 0.5694444     |
| Average          | 0.3905212     |

The calculated numbers during the period prior to 1875 clearly vary from one newspaper to another, with the smallest being 0 and the largest being 0.5970985. The average value indicates that, in general, the newspapers belonged to few groups. How do we check the statistical significance of this result? To do this, I ran a simulation experiment in which I pooled the newspaper issues prior to 1875 together and randomly assigned them to 14 groups. The only condition I introduced was to keep the group sizes equal to the sizes of the groups obtained via the cluster analysis. I then calculated the Simpson index for each newspaper and the overall average. I

Table 4. Simpson index for the newspaper titles after 1874.

| Newspaper  | Simpson index |
|------------|---------------|
| Basheer    | 0.3223178     |
| Janna      | 0.00          |
| Jinan      | 0.415         |
| Lisan      | 0.1399177     |
| Mishkat    | 0.00          |
| Moktataf   | 0.00          |
| Nashra     | 0.00          |
| Tabeeb     | 0.00          |
| Taqadom    | 0.00          |
| Thamarat   | 0.1014031     |
| Average    | 0.0978638     |
repeated this trial 10,000 times, each time returning the average of the Simpson indexes obtained. This experiment allowed me to obtain the expected values of the Simpson index if the assignment of the issues was in fact random. The result was that the 98% confidence interval of the returned value was [0.5761655, 0.685548]. Table 3 shows a value of 0.3905212, which is considerably outside the obtained interval. What does this mean? Since 0.3905212 is less than the lower limit of the 98% confidence interval, it means that our obtained average from the Lebanese industry is in fact not due to randomness. The newspapers in this early period were in fact producing issues that, to some extent, seemed to fall into a small number of groups.

With regard to the years 1875—1879, Table 4 clearly shows that there was a decrease in the number of groups spanned by each newspaper. In fact, the most frequently occurring number in the table is zero. This indicates that, for the most part, all issues of each newspaper fell into a single group in our cluster analysis. Again, to study the statistical significance of the number obtained, I ran a similar simulation to the one described above, but this time on the period 1875—1879. The 98% confidence interval of the average of the Simpson index I obtained was [0.3946162, 0.5367033]. Had the issues been randomly assigned to the five groups, then we would have expected the average of the Simpson index to lie within that interval. However, the value we obtained was 0.0978638, which is much less than the lower limit of the 98% confidence interval. This tells us that the issues belonging to the same newspapers during that period mostly fell within the same group and that this was not due to randomness. In addition, the number obtained is much less than that obtained for the period 1851—1874, so it seems that later newspaper issues of each title were more likely to be in a smaller number of groups than those published before 1875. Again, the question arises as to the statistical significance of this decrease in the average of the Simpson indexes. What if this decrease was due to the fact that we divided the issues into a smaller number of groups (5 as opposed to the previous 14). Surely, this alone would result in a smaller Simpson index. Once again, I used a simulation experiment, but this time, I pooled all issues from 1851 to 1879 together and then randomly assigned them into one of two groups. The only constraint imposed was that the two groups be equal in size to the groups in my original dataset. After assigning the issues to the two groups, I re-ran the simulations described above on both groups, but this time, I was interested in the difference between the two averages of the Simpson index. The 98% confidence interval of the difference obtained was [0.0160145, 0.160887]. The difference in the original dataset is 0.3905212−0.0978638 = 0.2926574, which is larger than the upper bound, with a p-value that is significant at the 0.001 level. The fact that it is larger than the upper limit of the confidence interval suggests that there is a larger than expected decrease in the Simpson index. This means that the issues after 1874 formed fewer groups and that the issues relating to the individual newspapers were more likely to end up in the same group than the 1851—1879 period. This lends support to Hypothesis 1.
The above analysis studied how the issues spanned the groups found in the cluster analysis. A question remains regarding the spanning of the different categories (politics, literature, economy, etc.). If a newspaper title has all its issues end up in a single group in the cluster analysis, then this does not mean that the newspaper’s contents dealt with just one category. The group formed in the cluster analysis may be composed of newspapers that spanned the same categories in the same manner. Above, we found that the newspapers in the period 1875—1879 spanned fewer groups, but did they span fewer categories? To study this question, I again calculated the Simpson index, but this time, I based the calculations on the ratio of the content of each category and not on the number of issues in each group. Instead of calculating the index for a newspaper title by dividing the total number of issues in a certain group by the total number of newspapers, I now calculated the index for each individual index by dividing the total amount of space dedicated to politics, for example, by the total space available in the issue. In this way, the index will tell us the niche of the issue with regard to the nine categories. After calculating the indexes, I calculated the average of all the issues in each group of the cluster analysis section. In this way, I could see whether a group was composed of specialist newspapers or more-general newspapers. Tables 5 and 6 show the results obtained for both periods.

If we concentrate on the largest two groups in both tables, we notice that the lowest average in the period 1875—1879 is smaller than the lowest average in the 1851—1874 period. We also notice that the reverse is true for the higher of the two averages in both tables, that is, it is greater in the period 1851—1874 than the period 1875—1879. The tables seem to suggest that the issues of the period 1875—1879 were more specialized in their content than those of the 1851—1879 period. If this was really the case, then not only did newspapers span a smaller number of group clusters in the latter period, but they also spanned a smaller number of categories. This would suggest that with time, newspapers became more alike and more focused in their content.

Table 5. Average Simpson index for the cluster groups of 1851—1874.

| Cluster group | Number of issues in the group | Average of the Simpson index |
|---------------|-------------------------------|-------------------------------|
| 1             | 108                           | 0.378203                      |
| 2             | 8                             | 0.6112977                     |
| 3             | 4                             | 0.5486086                     |
| 4             | 13                            | 0.5460258                     |
| 5             | 2                             | 0.5330223                     |
| 6             | 15                            | 0.5907483                     |
| 7             | 167                           | 0.5635558                     |
| 8             | 2                             | 0.7077688                     |
| 9             | 10                            | 0.6905794                     |
| 10            | 6                             | 0.2221736                     |
The final analysis concerns the number of clusters that best fit the newspapers published after 1874. Do the data suggest that all, or most, of these issues can be included in one single cluster, thereby giving support to the argument that organizations need to project a unified identity (McKendrick and Carroll, 2001; McKendrick et al., 2003), or do the newspapers cluster into more than one group, hence supporting the finding that more than a single identity is projected by organizations in an emerging industry (King et al., 2011)? I will use both the Calinski and Harabasz pseudo-F index and the Duda-Hart index to determine the optimal number of clusters for the given dataset. Table 7 shows the results of both measures calculated for 1 up to 10 clusters. With regard to the Calinski and Harabasz measure, the higher the value, the better the fit. Unfortunately, this measure gives us no information with regard to having a single cluster, but our second measure does. With regard to the Duda-Hart index, the best combination is the one that has as high a value as possible for the Je(2)/Je(1) term and as low a value as possible for the pseudo T-squared. The best values for both measures are in indicated in bold in Table 7.

Our first measure indicates that the number of clusters that best fit the data is two, while the second measure indicates that the number is three. If the data are divided

| Number of clusters | Calinski and Harabasz index | Duda-Hart index |
|--------------------|-----------------------------|----------------|
|                    |                             | Je(2)/Je(1)    | Pseudo T-squared |
| 1                  | —                           | 0.3013         | 684.04           |
| 2                  | **684.04**                  | 0.8128         | 33.85            |
| 3                  | 380.97                      | **0.9825**     | **2.43**         |
| 4                  | 255.24                      | 0.9752         | 3.72             |
| 5                  | 195.08                      | 0.8810         | 18.23            |
| 6                  | 164.66                      | 0.1794         | 663.28           |
| 7                  | 361.59                      | 0.5896         | 6.26             |
| 8                  | 313.83                      | 0.5778         | 70.89            |
| 9                  | 295.96                      | 0.9137         | 4.35             |
| 10                 | 265.19                      | 0.6732         | 62.62            |

The numbers in bold represent the best value of the Calinski and Harabasz index and of the Duda-Hart index respectively.
into two groups, based on the findings of the first statistic, then the distribution of the issues in the three groups will be as follows: one group with 149 issues and a second group with 148 issues. If, on the other hand, the data are divided into 3 groups, as the second measure suggests, then the distribution of the issues in the four groups will be as follows: one group with 148 issues, a second group with 138 issues, and a third group with 11 issues. In both cases, we see the same pattern, which is two large and equal-sized groups and a very small number of residual issues that do not seem to fit into either group. This result suggests that no single cluster emerges from the dataset. Instead, the issues seem to form two distinct groups of equal size, thus supporting Hypothesis 3B.

Next, I take a closer look at the contents of these two clusters in order to see the extent to which each cluster spans categories. Table 8 shows the averages of the ratios for each category in both clusters. The general finding is that we have two types of newspapers. The first has more than half its content dedicated to politics, and the second has most of its content dedicated to knowledge. Thus, by the end of the 1870s, there were two clear identities: political and scientific newspapers.

I conducted the same analysis on the issues published before 1875. In the first part of this paper, we found that issues published before 1875 were less similar to each other than issues published after 1874. Based on this, we would expect that these early issues would have a more difficult time clustering together. In other words, we would expect to see more than just two clusters forming. To test whether this was actually the case, I calculated the same statistics for the period 1851–1874 that were described above. Table 9 shows the results obtained.

Here, we clearly see that both measures differ significantly with regard to the optimal number of clusters that fit the data. This is not a strange finding in cluster analysis. Although both measures give different numbers of clusters, what concerns us here is that in both cases, the statistics point to the fact that more clusters are formed by

**Table 8.** The averages of the categories in the largest two clusters.

| Category    | Cluster 1 | Cluster 2 |
|-------------|-----------|-----------|
| Politics    | 0.1051542 | 0.6378778 |
| Economics   | 0.0060576 | 0.0313665 |
| Social      | 0.0113373 | 0.0493177 |
| Literature  | 0.0639691 | 0.0511335 |
| Advertisement | 0.0023351 | 0.0211913 |
| Knowledge   | 0.7814186 | 0.0801641 |
| Art         | 0.0003097 | 0.000906  |
| Sport       | 0         | 0.000168  |
| Other       | 0.0252021 | 0.1283396 |
Table 9. Results for finding the optimal number of clusters 1851—1874.

| Number of clusters | Calinski and Harabasz index | Duda-Hart index |
|--------------------|----------------------------|-----------------|
|                    |                           | Je(2)/Je(1)     | Pseudo T-squared |
| 1                  | —                          | 0.9707          | 10.87            |
| 2                  | 10.87                      | 0.6012          | 232.85           |
| 3                  | 125.68                     | 0.7117          | 59.96            |
| 4                  | 122.61                     | 0.6960          | 57.23            |
| 5                  | 120.23                     | 0.8005          | 50.08            |
| 6                  | 117.80                     | 0.9065          | 18.25            |
| 7                  | 104.75                     | 0.5389          | 5.99             |
| 8                  | 91.64                      | **0.9829**      | **2.91**         |
| 9                  | 80.85                      | 0.6471          | 2.73             |
| 10                 | 72.14                      | 0.8749          | 16.87            |

The numbers in bold represent the best value of the Calinski and Harabasz index and of the Duda-Hart index respectively.

these issues than the number formed by issues published between 1875 and 1879\(^2\). We can see that the scores of the Calinski and Harabasz index for three, four, five, and six groups are close to each other and are much larger than the score for two clusters. In fact, the score for two clusters is the lowest of all the scores. The Duda-Hart index indicates that the issues are best grouped into eight clusters, lending support to Hypothesis 2. This finding shows that issues published before 1875 form a larger number (Hypothesis 2) of less-coherent clusters (Hypothesis 1) than the issues published after 1874.

6.2. Multilevel analysis

The next step in our analysis is to try to explain the change in the category spanning dynamics and to see whether the data support Hypothesis 4. In our case, I thought of three events of interest. The first dependent variable, called Exiting, would record an event in which a current newspaper issue does not span a category that it used to span in the previous issue. The second dependent variable, called Entering, would record the event of spanning a category that it had not spanned in the previous issue. Finally, I created the third dependent variable by using the logical OR operator. This third variable takes on the value of one when either of the previous two dependent variables are one. The reason for the creation of this variable is that while each of the first two dependent variables tracked a dynamic that had a certain direction (either

---

2. To test the robustness of the results, I repeated the analysis using median linkage instead of weighted average linkage. The results indicated that the period 1875—1879 produced fewer clusters than the period 1851—1874.
exiting or entering), the third variable has no direction. This third variable measures the unsystematic dynamic because no matter what direction the newspaper takes, this variable will record a one with any change. An important issue here is the following: when do we consider that a newspaper has spanned a category? In the dataset, there are instances in which one newspaper issue does not dedicate any space to literature, but in the next issue, the same newspaper may dedicate only 5% to it. Should this be considered a case of spanning? Probably not. So what is the lower limit that should be accepted? Since there are nine categories in total, I took 0.11 (1/9) as the accepted limit. In a random scenario, an issue would dedicate 0.11 of its space to every single category. Therefore, when a newspaper dedicates more than that amount to a category, I considered it as spanning that category. These three variables are recorded as a series of ones and zeros and are hence modeled using logistic regression (Hilbe, 2009).

Because the data are of a longitudinal nature, and because observations belonging to the same newspaper will probably not be independent, I use a random effects model. More specifically, I include a newspaper-specific random-intercept $\zeta_j$ in order to relax the assumption of the independence of observations that pertain to the same newspaper, as follows (Rabe-Hesketh and Skrondal, 2012):

$$\logit \{ \Pr(y_{ij} = 1| x_{ij}, \zeta_j) \} = \beta_0 + x\beta + \zeta_j$$

where the subscript $j$ refers to the newspaper, while the subscript $i$ refers to the individual observation within each panel. I use an identity covariance structure because the best-fit model will include only a random-intercept, and the identity covariance structure is the only possible structure in such a case (Stata Corporation, 2011).

There are four control variables in the base model. The first is the type of publication (newspaper or magazine). This controls for the fact that a certain type of publication might be more inclined to span categories, and an increase in the number of such a publication in a certain period might be the cause of the increase in the probability of spanning categories. The second variable is the frequency of publication of the issues. Perhaps newspapers with a higher frequency would have a higher chance of spanning categories. The third variable is the number of pages of the issue. More pages might mean a higher probability of category spanning. The final variable is the total number of newspapers active during the time of publication. More titles might also cause an increase in category spanning.

The results of this analysis are shown in Table 10. First we start with the dependent variable Exiting. We see from the base model that issues from bi-weekly newspapers...
Table 10. Multi-level logistic regression of the three variables Exiting, Entering, and Unsystematic movement.

|                      | Exiting categories |                      | Entering categories |                      | Unsystematic category movement |
|----------------------|--------------------|----------------------|--------------------|----------------------|-------------------------------|
|                      | Base               | Model 1              | Model 2            | Model 3              | Base                          | Model 1              | Model 2              | Model 3              |
| Coef.                | Coef.              | Coef.                | Coef.              | Coef.                | Coef.                         | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                | Coef.                |
| Frequency (base is bi-monthly) |                   |                      |                    |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                      |                               |                               |                               |
| Bi-weekly            | 1.701343*          | 1.562821*            | 1.604934           | 1.095462             | 1.103122                     | .902594                        | 1.952803             | 1.730702             | 1.773637             | 1.402561                       |                               |                      |                    |                      |                      |                               |                               |                               |
| Irregular            | .3156745           | .3280339             | -.0951772          | -.2296452            | -.2684769                    | -.2993611                     | -.4575896            | -.5349552            | -.6924957            | -.7481581                       | -.1602802                   |                   | -.1151953           |                      |                      |                               |                               |                               |
| Monthly              | -.0344205          | .0967091             | .0308954           | -.310834             | -.0769286                    | -.0404292                     | -.0516032            | -.2525044            | -.2590298            | -.1124904                       | -.184569                     | -.643778           |                      |                      |                      |                               |                               |                               |
| Weekly               | .9717399           | .851204              | .7944905           | .7178168             | .6234923                     | .5662044                      | .5382138            | .4804657             | 1.046017             | .868288                          | .7073139                     | .8029872          |                      |                      |                      |                               |                               |                               |
| Yearly               | -.17.80174         | -.18.61402           | -.17.8258          | -.18.24017           | -.19.5504                     | -.1.138448                    | -.1.475437           | -.1.231831           | -.2.713508           | -.2.471926                       | -.4.060184*                    | -.2.671821         |                      |                      |                      |                               |                               |                               |
| Type (base is magazine) |                   |                      |                    |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                               |                               |                               |
| newspaper            | .34832             | .6267084             | .3637016           | .3420479             | .238569                      | .3086556                      | .2479894            | .2645217             | -.1.805581           | -.0558052                        | -.1.231385                    | -.0826729       |                      |                      |                      |                               |                               |                               |
| Pages                | .0047411           | .0054431             | -.0002743          | -.0014483             | -.0112505                    | -.0118452                     | -.0134791          | -.0124534            | -.04170725          | -.0190181                       | -.02498971                    | -.0015326       |                      |                      |                      |                               |                               |                               |
| Active newspapers    | -.1.327501*        | -.0773368            | -.0676776          | -.1.304601**          | -.1.443052**                  | -.1.291721*                   | -.1.17652           | -.1.351247**         | -.2.261044**         | -.1.687694                        | -.1.253563                    | -.1.928283** |                      |                      |                      |                               |                               |                               |
| Period               | -.418046           | -.140387             | -.4900159          |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                               |                               |                               |
| Time                 | -.555299           | -.0234373            | -.1.035357*        |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                               |                               |                               |
| Age                  | -.0068259*         | -.0043614            | -.0102867**        |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                               |                               |                               |
| Constant             | -.4032355          | -.805512             | -.3946393          | -.1.727559            | -.3.297689                    | -.2.608948                    | -.6.810376          | -.5.688876           | -.2.092221           | -.1.795075                        | -.3.736369*                    | -.2.420477     |                      |                      |                      |                               |                               |                               |
| ln sig2u             |                   |                      |                    |                      |                               |                               |                      |                    |                      |                               |                               |                      |                    |                      |                      |                               |                               |                               |
| Constant             | -.2.340204         | -.3.195689           | -.2.112843         | -.2.361705*            | -.1.494508                    | -.1.566694                    | -.1.440816          | -.1.590217           | -.4.350034           | -.5.545673                        | -.3.505714                    | -.5.528825     |                      |                      |                      |                               |                               |                               |
| AIC                  | 770.0822           | 768.8263             | 770.0582           | 767.3396              | 775.6501                     | 776.4511                      | 777.2981            | 775.7795             | 722.227             | 721.5815                          | 718.6239                      | 715.9555      |                      |                      |                      |                               |                               |                               |
| BIC                  | 814.7125           | 817.9196             | 819.1515           | 816.4329              | 820.2804                     | 825.5444                      | 826.3914            | 824.8728             | 766.8573             | 770.6748                          | 767.7173                      | 765.0489      |                      |                      |                      |                               |                               |                               |
| likelihood           | -.375.0411          | -.3.373.4131         | -.3.374.0291       | -.3.372.6698           | -.3.377.8251                   | -.3.377.2256                   | -.3.377.649        | -.3.376.8988          | -.351.1135           | -.3.349.7908                        | -.3.348.312         | -.3.346.9778 |                      |                      |                      |                               |                               |                               |
| N. of cases          | 641                | 641                  | 641                | 641                   | 641                           | 641                           | 641                 | 641                   | 641                  | 641                               | 641                            | 641              |                      |                      |                      |                               |                               |                               |

*p < 0.05, **p < 0.01, ***p < 0.001.
are statistically more likely to exit categories than other publications with different frequencies. Only the coefficient of bi-weekly is statistically significant. The only other variable that is also statistically significant is the number of active newspapers, which has a negative coefficient. I next created a period indicator variable that was set to zero prior to 1875 and to one for issues published in the second period. This variable is used to test whether the issues of one period exited categories more often than the issues of the other period. The results (under Model 1) show that the coefficient is negative but not statistically significant. In Model 2, I replaced the period indicator with a continuous variable that recorded the time since the emergence of the population (the date of the very first newspaper). Again, the coefficient is negative but not statistically significant. Finally, in Model 3, I replaced the time variable with the age of the newspaper. Because the time, both continuous and binary, was not significant, perhaps the age of the individual newspapers was a better explanatory variable. We see that the coefficient of the variable age is negative and statistically significant. This means that the older the newspaper became, the less likely it was to exit a category. Next, I followed the same strategy with regard to the second and third dependent variables. With regard to the variable *Entering*, the three added time variables were negative and statistically not significant. Thus, it seems that while newspapers tended to decrease their category exiting activities with age, they retained their category entering activities.

Finally, with regard to the last dependent variable, we see that both time and age are negative and statistically significant, but the model with age produces lower AIC and BIC statistics, thus indicating that age was the better predictor.

What do these results tell us? We have seen from Fig. 3 that during the initial years, there was a large variation in the average monthly Simpson index. This variation decreased considerably in later years. The regression results shown above inform us that as newspapers aged, they were less likely to exit categories and were less likely to follow a random pattern of entering and exiting. With regard to entering categories, we saw that none of the time variables were statistically significant. Together, these results show that the founding newspapers moved unsystematically
between categories and that with age, they ceased exiting categories but continued their entering dynamics, thus lending support to Hypothesis 4. In fact, we can see from Fig. 3 that during the second period (after 1874), the average Simpson index started rising. This can be attributed to the fact that the newspapers stopped the exiting dynamic and continued the entering dynamic.

7. Conclusion

This paper has used the logical formulations introduced by Hannan et al. (2007) to take an in-depth look at both the group-spanning and category-spanning dynamics of the Lebanese newspapers at the time of their emergence. Unlike previous research that determined the categories organizations spanned by referring to third parties such as critics (Zuckerman, 2000) or to official organizational announcements and publications (Pontikes, 2008), this research extracted the information from the actual product itself. This information was used to study two levels of spanning. The first was the spanning that took place on the level of the categories themselves, while the second was the spanning that took place on the level of the clusters that were formed by the data. In both cases, it was found that during later years in the formation period, the newspapers became much more consistent in their data coverage in their respective issues, and they became more similar to each other. This result was supported by both cluster analysis and multi-level logistic regression.

Table 11 shows the previous studies and the findings of each study. This table is presented to put the results of my study in perspective and to see how it fits into the picture. As can be seen from the table, no consensus has been found regarding the issue of whether a unified identity is a necessary condition for successful emergence. This study supports the view that a unified identity is not a necessary condition, as do three of the other studies.

Another important result was that the issues of the first period (1851—1874) did not produce haphazard content by jumping from category to category. Ever since the inception of the industry, the individual issues of the newspapers showed a tendency to display similar content, a tendency that increased with time. Another important finding was that in the middle of the 1870s, there was a change in the content of the newspapers. Interestingly, this change happened around the same time that a change in the goals and motivation of the newspapers took place, as argued by Mozahem (2015). The observed change is best described as a process of maturation with age. The older the newspapers became, the more focused their content became. Finally, this study found that two clearly separate identities had formed by the end of 1870s. One represented political newspapers, and the other represented scientific newspapers. It took a significant amount of time for these two identities to be clearly separated. Therefore, this study has shown that, at least with regard to that part of the identity that is determined by the categories spanned by the organization, industries can and do emerge by projecting more than a single identity.
Declarations

Author contribution statement

Najib Mozahem: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Albert, Stuart, Whetten, David A., 1985. In: Cummings, L.L., Staw, B.M. (Eds.), Research in Organizational Behavior. JAI Press, Greenwich, pp. 263–295.

Aldrich, Howard, Ruef, Martin, 2006. Organizations Evolving. Sage, California.

Beck, Nikolaus, Swaminathan, Anand, Wade, James B., Wezel, Filippo Carlo, 2009. The Impact of Organizational Prototypes and Industry Clusters on the Fates of Northern Bavarian (Franconian) Breweries. Retrieved from https://business.illinois.edu/business-administration/wp-content/uploads/sites/39/2014/09/wade_paper.pdf.

Berger, Peter L., Luckmann, Thomas, 1967. The Social Construction of Reality: a Treatise in the Sociology of Knowledge. Anchor Books, New York.

Biggart, Nicole Woolsey, Beamish, Thomas D., 2003. The economic sociology of conventions: habit, custom, practice, and routine in market order. Annu. Rev. Sociol. 29, 443–464.

Carroll, Glenn R., Swaminathan, Anand, 2000. Why the microbrewery movement? Organizational dynamics of resource partitioning in the U.S. Brewing industry. Am. J. Sociol. 106, 715–762.

Carroll, Glenn R., Bigelow, Lyda S., Seidel, Marc-David L., Tsai, Lucia B., 1996. The fates of de novo and de alio producers in the american automobile industry 1885–1981. Strat. Manag. J. 17, 117–137.
Carroll, Glenn R., Hannan, Michael T., 2004. The Demography of Corporations and Industries. Princeton University Press.

Everitt, Brian S., Landau, Sabine, Leese, Morven, Stahl, Daniel, 2011. Cluster Analysis. Wiley.

Hamilton, Lawrence C., 2009. Statistics with Stata: updated for Version 10. Brooks/Cole.

Hannan, Michael T., 2010. Partiality of memberships in categories and audiences. Annu. Rev. Sociol. 36, 159–181.

Hannan, Michael T., Baron, James N., Hsu, Greta, Koçak, Özgecan, 2006. Organizational identities and the hazard of change. Ind. Corp. Change 15, 755–784.

Hannan, Michael T., Pólos, László, Carroll, Glenn R., 2007. Logics of Organization Theory: Audiences, Codes, and Ecologies. Princeton University Press.

Hilbe, Joseph M., 2009. Logistic Regression Models. CRC Press.

Hsu, Greta, Hannan, Michael T., 2005. Identities, genres, and organizational forms. Organ. Sci. 16, 474–490.

Hsu, Greta, Hannan, Michael T., Koçak, Özgecan, 2009a. Multiple category memberships in markets: an integrative theory and two empirical tests. Am. Socio. Rev. 74, 150–169.

Hsu, Greta, Hannan, Michael T., Pólos, László, 2009b. Typecasting and Legitimation: a Formal Theory. Stanford Graduate School of Business.

Hsu, G., Hannan, M.T., Pólos, L., 2011. Typecasting, legitimation, and form emergence: a formal theory. Socio. Theor. 29, 97–123.

Jensen, Michael, 2010. Legitizing illegitimacy: how creating market identity legitimizes illegitimate products. In: Hsu, G., Negro, G., Kocak, O. (Eds.), Categories in Markets: Origins and Evolution, Research in the Sociology of Organizations. Emerald Group Publishing Limited, UK, pp. 39–80.

Khaire, Mukti, Wadhwani, R. Daniel, 2010. Changing landscapes: the construction of meaning and value in a new market category - modern indian art. Acad. Manag. J. 53, 1281–1304.

Khessina, Olga M., Carroll, Glenn R., 2008. Product demography of de novo and de alio firms in the optical disk drive industry, 1983–1999. Organ. Sci. 19, 25–38.

King, Brayden G., Clemens, Elisabeth S., Fry, Melissa, 2011. Identity realization and organizational forms: differentiation and consolidation of identities among arizona’s charter schools. Organ. Sci. 22, 554–572.
Kocak, Ozgecan, Hannan, Michael T., Hsu, Greta, 2009. Audience structure and category systems in markets. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.622.4012.

Kovacs, Balazs, Hannan, Michael T., 2009. Category Contrast, Spanning, and Appeal: Food-service Organizations. Retrieved from http://media.coauthors.net/konferencia/conferences/l/kovacs%26hannan_oct5.pdf.

Krippendorff, Klaus, 2004. Content Analysis: an Introduction to its Methodology. Sage.

Lamont, Michèle, Molnár, Virág, 2002. The study of boundaries in the social sciences. Annu. Rev. Sociol. 28, 167–195.

Liu, Min, Wezel, Filippo C., 2013. Small Is Beautiful? Organizational Identity and Growth Rates in a Partitioned Market. Working Paper. Durham Research Online (DRO), Durham. http://dro.dur.ac.uk/11414.

Lounsbury, Michael, Rao, Hayagreeva, 2004. Sources of durability and change in market classifications: a study of the reconstitution of product categories in the american mutual fund industry, 1944–1985. Soc. Forces 82, 969–999.

McKendrick, David G., Carroll, Glenn R., 2001. On the genesis of organizational forms: evidence from the market for disk arrays. Organ. Sci. 12, 661–682.

McKendrick, David G., Jaffee, Jonathan, Carroll, Glenn R., Khessina, Olga M., 2003. In the bud? Disk array producers as a (possibly) emergent organizational form. Adm. Sci. Q. 48, 60–93.

Milligan, Glenn W., Cooper, Martha C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Mozahem, Najib A., 2017. Cluster formation as a representation of the category space: a two-level theoretical model tested within the context of the lebanese newspaper industry (1851-1974). SAGE Open. April–June: 1–16.

Mozahem, Najib A., 2015. Identities, Categories, and Clusters: a Study of Category Dynamics and Cluster Spanning in the Lebanese Newspaper Industry 1851-1974. PhD Thesis. Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/11348/.

Navis, Chad, Glynn, Mary Ann, 2010. How new market categories emerge: temporal dynamics of legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005. Adm. Sci. Q. 55, 439–471.

Negro, Giacomo, Kocak, Ozgecan, Hsu, Greta, 2010. Research on categories in the sociology of organizations. In: Hsu, G., Kocak, O., Negro, G. (Eds.), Categories in
Markets: Origins and Evolution, Research in the Sociology of Organizations. Emerald Group Publishing Limited, UK, pp. 3–38.

Perretti, Fabrizio, Negro, Giacomo, Lomi, Alessandro, 2008. E pluribus unum: framing, matching, and form emergence in U.S. Television broadcasting, 1940–1960. Organ. Sci. 19, 533–547.

Pólos, László, Hannan, Michael T., 2001. Nonmonotonicity in theory building with applications to organizational mortality. In: Lomi, A., Larsen, E. (Eds.), Dynamics of Organizations: Computational Modeling and Organizational Theories. MIT Press, Massachusetts, pp. 405–437.

Pólos, László, Hannan, Michael T., 2002. Reasoning with partial knowledge. Socio. Meth. 32, 133–181.

Pólos, László, Hannan, Michael T., Carroll, Glenn R., 2002. Foundations of a theory of social forms. Ind. Corp. Change 11, 85–115.

Pontikes, Elizabeth G., 2008. Fitting in or Starting New? an Analysis of Invention, Constraint, and the Emergence of New Categories in the Software Industry. PhD Thesis. Stanford Graduate School of Business, Stanford, CA.

Powell, Walter W., Sandholtz, Kurt W., 2011. Amphibious entrepreneurs and the emergence of new organizational forms. Strat. Entrepren. J. 6, 94–115.

Pozner, Jo Ellen, Rao, Hayagreeva, 2006. Fighting a common foe: enmity, identity and collective strategy. Adv. Strat. Manag. 23, 445–479.

Rabe-Hesketh, Sophia, Skrondal, Andres, 2012. Multilevel and Longitudinal Modeling Using Stata Vol. 2: Categorical Responses, Counts, and Survival. Stata Press.

Riffe, Daniel, Lacy, Stephen, Fico, Frederick, 1998. Analyzing Media Messages: Using Quantitative Content Analysis in Research. L. Erlbaum.

Ruef, Martin, 2000. The emergence of organizational forms: a community ecology approach. Am. J. Sociol. 106, 658–714.

Ruef, Martin, Patterson, Kelly, 2009. Credit and classification: the impact of industry boundaries in nineteenth-century America. Adm. Sci. Q. 54, 486–520.

Santos, Filipe, Eisenhardt, Kathleen, 2009. Constructing markets and shaping boundaries: entrepreneurial power in nascent fields. Acad. Manag. J. 52, 643–671.

Smith, Edward Bishop, 2011. Identities as lenses: how organizational identity affects audiences’ evaluation of organizational performance. Adm. Sci. Q. 56, 61–94.
Stata Corporation, 2011. Stata Longitudinal-data/panel-data Reference Manual: Release 12. Stata Press.

Weber, Klaus, Heinze, Kathryn L., DeSoucey, Michaela, 2008. Forage for thought: mobilizing codes in the movement for grass-fed meat and dairy products. Adm. Sci. Q. 53, 529–567.

Wei, Zhao, 2005. Understanding classifications: empirical evidence from the American and French wine industries. Poetics 33, 179–200.

Wry, Tyler, Lounsbury, Michael, Glynn, Mary Ann, 2011. Legitimating nascent collective identities: coordinating cultural entrepreneurship. Organ. Sci. 22, 449–463.

Zuckerman, Ezra W., 2000. Focusing the corporate product: securities analysts and de-diversification. Adm. Sci. Q. 45, 591–619.