CDF/PHYSICS/TOP/2023
April 1, 1993

B Identification Using Jet Probability

David Gerdes
University of Michigan

1 Introduction

With run Ia now past its midpoint and over 16 pb^{-1} of data on tape, several facts have become clear concerning the top search. First, the presence of only one bona fide e-μ candidate[1] and only a small handful of b-tagged lepton+jets events[2] indicates that top is quite heavy. As a consequence, it becomes a matter of great importance to achieve the highest possible b-tagging efficiency. Second, the low rates for the top signal demand a very careful analysis of the backgrounds, and any b-tagging algorithm must lend itself readily to such a study. Finally, since any discovery claim will rest on possibly as few as half a dozen events, treating the analysis solely as a counting experiment with x tagged events on a background of y may not be sufficient—we will want to make specific, quantitative statements about individual jets and even individual tracks, statements more quantitative and more convincing than simply whether such an object either passed or did not pass a particular tag.

This note describes a b-tagging algorithm based on a jet probability function. The function is calculated by comparing the the tracks in the jet with positive signed impact parameters to the measured SVX resolution function. This probability distribution is flat for zero-lifetime jets and sharply peaked at zero for b-jets; physically it represents the probability that the observed set of impact parameters is consistent with resolution effects alone. Charm and strange particles, which also have a lifetime but which fragment into lower-multiplicity jets, produce a much less sharply peaked distribution. One can therefore impose a b-tag by requiring the jet probability to be less than some value, say 1%.

There are many advantages to this approach to b-tagging. Because the jet probability is well-defined for almost every SVX-fiducial jet (it requires only one track with a positive impact parameter that satisfies minimal quality cuts), it is capable of very high efficiencies. The b-tagging algorithms that have been used to date—the jet-vertexing[3], cone-tag[4], and D-ϕ[5] algorithms—have measured efficiencies[3,6] that range from about 8 to 18% per SVX-fiducial jet. (Subsequently the Jetvtx efficiency has improved somewhat from its original 15% value owing to an improved treatment of shared hits.) The jet probability tag easily achieves 20% efficiency with low background rates, and even if the jet probability cut is relaxed to allow

30% efficiency or higher, the backgrounds may be acceptably low. Indeed, a very nice feature of this algorithm is that the b-tag is performed using a continuous variable, not a discrete object like a reconstructed secondary vertex or a D-ϕ cluster. It therefore provides a knob that allows one to move smoothly along the efficiency curve and sit at the optimal signal-to-background point for the analysis in question. Furthermore, the ability to continuously relax the jet probability cut is a valuable tool for the understanding of backgrounds. An additional tool is the jet probability formed from the *negative* signed impact parameter tracks in the jet. This will be discussed further below. Finally, the method is conceptually simple and gives distributions that are visually appealing.

The jet probability algorithm was originally developed at Aleph by CDF alumnus Dave Brown[7], for use in studying hadronic decays of the Z^0 with Aleph's 3-D VDET. It proves, however, to be well-suited for b-tagging at the Tevatron.

2 Description of the Method

2.1 Signed Impact Parameter

Fig. 1 shows two tracks with identical impact parameters. Also shown is the axis of a jet to which these tracks are associated. One can give a sign to the impact parameter according to the angle between the jet axis and the track's point of closest approach to the primary vertex. If this angle is acute—that is, if the track comes from the same side of the event vertex as the jet does—we say that the impact parameter is positive. If this angle is obtuse, the track comes from the opposite side of the event vertex and is given a negative impact parameter. In the absence of lifetime effects, the impact parameter is determined solely by the SVX resolution function, and will be positive or negative with equal probability. Particles with a lifetime, however, such as b hadrons, will travel some distance along the jet direction before decaying, and their decay products will therefore preferentially populate the positive side of the impact parameter distribution. This is illustrated in Fig. 2, which shows the signed impact parameter distribution for tracks in a generic jet sample (left) and for the muon track in an inclusive 9 GeV CMU-CMP muon sample (right). The distribution for the jet data is quite symmetric, with a small excess on the positive side. The inclusive muon sample, however, is believed to contain approximately 40% b's, and exhibits a notable excess of positive impact parameter tracks. (These plots suggest that fitting the signed impact parameter distribution is a good way to actually estimate the b fraction, and indeed such a technique was used by us[6] to obtain the 40% number quoted above.) Thus, the negative side of the signed impact parameter distribution characterizes the resolution function, while the positive side characterizes both the resolution and lifetime effects. We will use the negative side of the signed impact parameter distribution to measure the SVX resolution function. In practice, to minimize contributions from badly measured tracks with a large impact parameter,

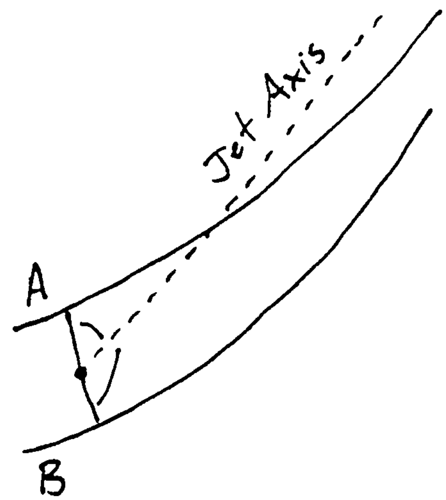

Figure 1: Two tracks with impact parameters equal in magnitude but opposite in sign. The jet direction is used to give a positive sign to track A, and a negative sign to track B.

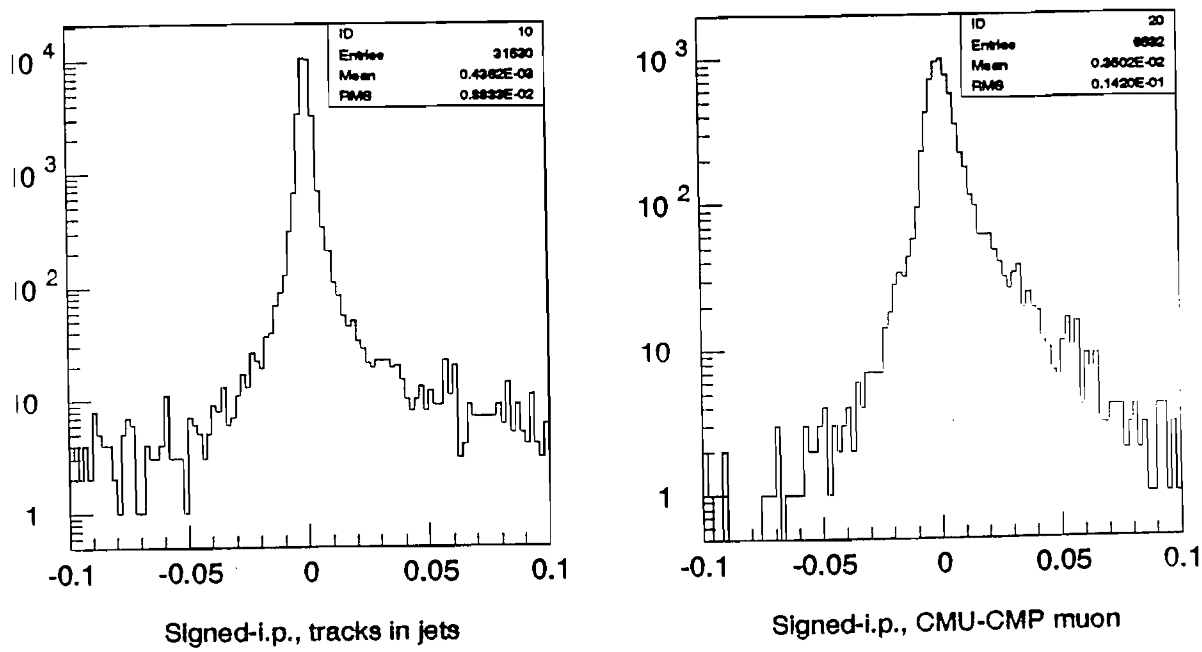

Figure 2: Signed impact parameter distributions for tracks in generic jets (left) and for the muon track in an inclusive CMU-CMP muon sample (right).

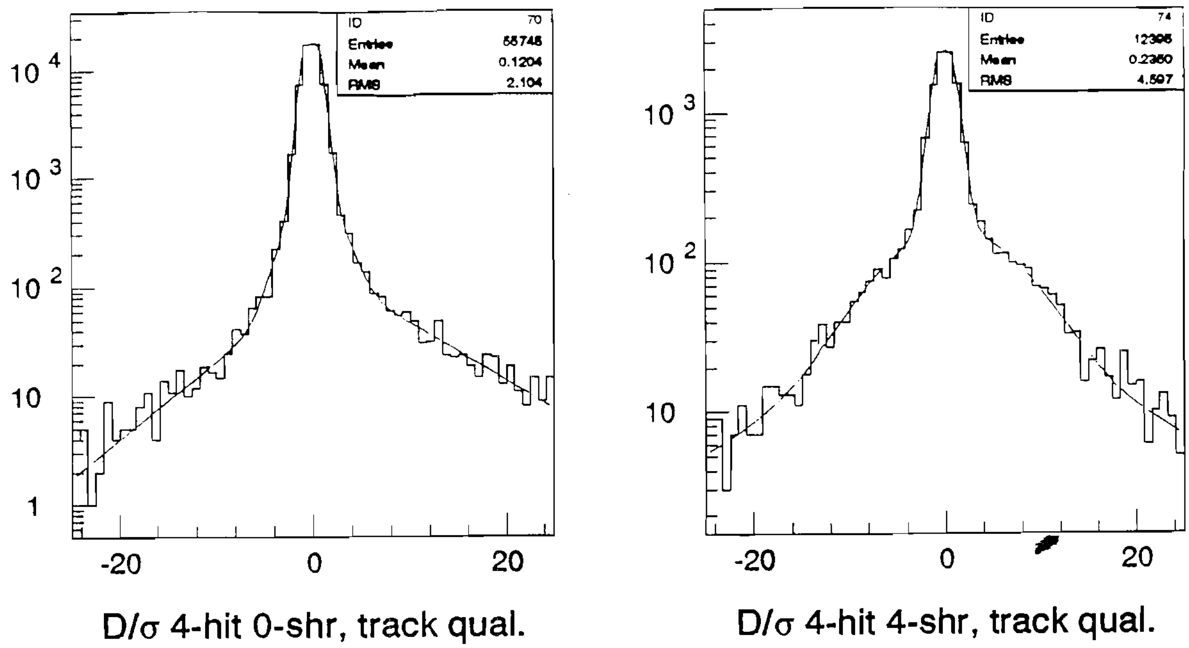

Figure 3: Signed impact parameter significance, s, for two categories of SVX tracks: 4-hit, 0-shared hits (left) and 4-hit, 4-shared hits (right).

it is better to use a related quantity, the signed impact parameter divided by its error, $s = D/\sigma$. The shape of this distribution is not the same for all types of tracks. Fig. 3 shows the s-distributions for 4-hit tracks with no shared hits, and for the extreme case of 4-hit tracks with 4 shared hits. Rather than reject such tracks outright, we simply regard them as defining a different, broader probability distribution; "large-s" tracks coming from such a distribution are therefore naturally de-weighted.

2.2 Track Probability and Measurement of the Resolution Function

To measure the resolution function, almost any large sample of tracks in jets is suitable; eventually for the top search one may want to the inclusive W and Z samples, thereby calibrating the algorithm on the analysis dataset itself. For the present study, however, we have used a large sample of Jet-50,70, and 100 triggers.[1] This sample was processed with Production version 6.01; the "bad-t_0" runs from December 1992 were excluded. We re-performed the SVX pattern-recognition and fit in a "6.1-like" way using the DO_SVX_PAD module, with the further refinement that the CTC errors were increased by a scale factor[8] of 2.6 to account for the overoptimism in 6.01. We then used tracks in jets with $E_T > 10$ GeV to measure the resolution function. Tracks were required to satisfy the following minimal cuts:

1. $P_T > 1$ GeV

[1]We are indebted to Brian Winer for making this sample available to us.

2. $\chi^2_{SVX} < 20$ (such a cut is efficient once the CTC errors have been rescaled)

3. absolute value of impact parameter, D, less than 0.1 cm

4. Primary vertex error from VXPRIM less than 60 microns.

The negative side of the s-distributions for each category of track—4-hit, 0-shared hit; 3-hit, 1-shared hit, etc.—defines a set of probability distributions that characterize the resolution function. We fit each such normalized distribution to a function $R(s; N_{hits}, N_{shr})$ (two such fits are shown in Fig. 3) a convenient functional form is two Gaussians plus an exponential tail. Then, we define the *track probability*

$$P_{track}(s; N_{hits}, N_{shr}) = \int_{-\infty}^{-|s|} R(t; N_{hits}, N_{shr})\, dt \tag{1}$$

This gives the probability that a track with N_{hits} hits and N_{shr} shared hits would have the observed value of s *or less*.

For tracks with a negative signed impact parameter, the distribution in $P(s)$ should be flat between 0 and 1, i.e. $P(s)$ should indeed behave as a probability, assuming we have parametrized the resolution function correctly. The actual distribution is shown in Fig. 4. It is essentially flat over most of the interval, but there is a small peak near zero. This indicates that there is a slight excess of tracks with a large negative s compared to the expectation from our fit. Possibly the tails fall more slowly than the exponential we have chosen; perhaps these very poor tracks could be removed by a suitable quality cut.

We expect lifetime effects to emerge when the same resolution function is used to calculate the track probability for positive signed impact parameter tracks. Fig. 5 shows this distribution, which indeed exhibits a pronounced peak near zero. The excess of low-probability (i.e. large-s) tracks indicates the presence of long-lived particles in the jet sample.

To cast the track probability distributions in a more familiar light, consider the case where the resolution function $R(s)$ is a Gaussian. The track probability distribution is then a complementary error function, $P(s) = \mathrm{erfc}(s)$, and a "1-σ track" has $P(s) = 0.317$. Similarly, a 2σ track has $P(s) = 0.046$, and a 3σ track has $P(s) = 0.0027$. Thus the simple cone-tag algorithm, which requires, say, at least three 3σ tracks, can be expressed in this language as requiring "three tracks with $P(s) \leq 0.0027$."

2.3 From Track Probability to Jet Probability

But making cuts on individual track probabilities is too restrictive, and does not exploit the full correlation of the tracks in B hadron decay. We choose instead to combine the probabilities of the positive signed impact parameter tracks in the jet into an overall *jet probability* equal to the probability that the jet would have tracks with the observed set of impact parameters *or any combination less probable.* To see

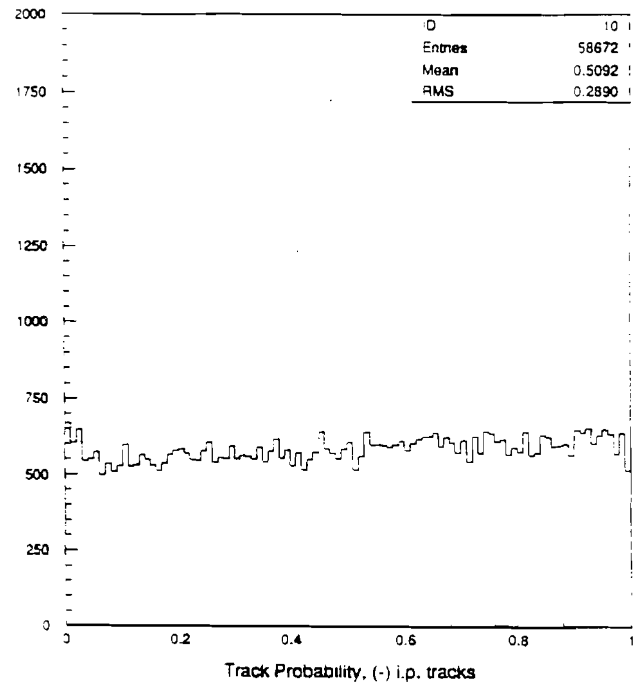

Figure 4: The track probability distribution for negative signed impact parameter tracks in the jet sample. These tracks were used to measure the SVX resolution function; the flatness of this distribution checks the accuracy of our fit.

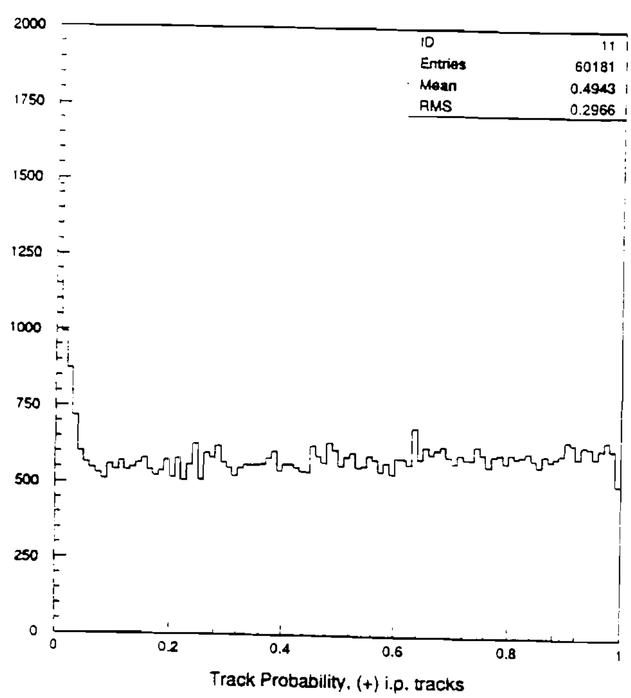

Figure 5: Track probability distribution for positive-signed impact parameter tracks in the jet sample. The excess near zero comes from particles with a lifetime.

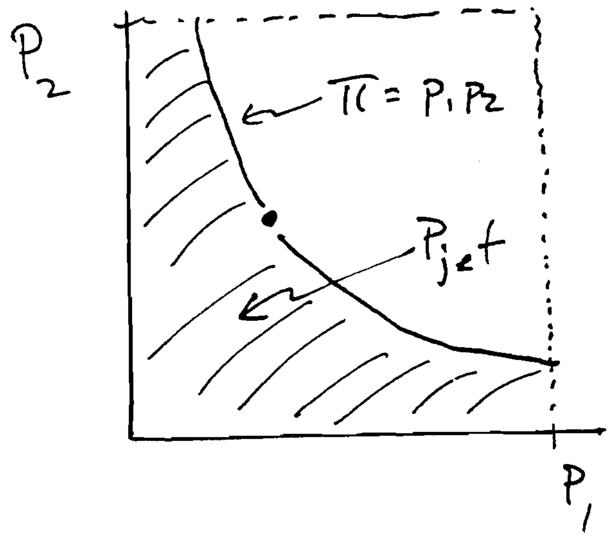

Figure 6: Definition of the jet probability for a two-track jet.

how this probability is constructed, consider a jet with two positive signed-i.p. tracks, with probability values P_1 and P_2 (see Fig. 6). Let $\Pi = P_1P_2$. Fig. 6 shows the curve of constant probability, and the area below and to the left of the curve is the set of two-track combinations with a probability less than or equal to Π. This area is defined to be the jet probability. For a two-track jet, the probability is $P_{jet} = \Pi(1 - \ln\Pi)$. In general, one can show inductively that

$$P_{jet} = \Pi \sum_{k=0}^{N-1} \frac{(-\ln\Pi)^k}{k!}, \tag{2}$$

where

$$\Pi = P_1P_2\cdots P_N \tag{3}$$

is the product of the individual probabilities of the positive-impact-parameter tracks.

The jet probability distribution for jets in the QCD sample is shown in Fig. 7. In this plot, and in all the ones that follow, I have excluded 4-hit tracks with 3 or more shared hits, 3-hit tracks with 2 or more shared hits, and 2-hit tracks with any shared hits. Jets are required to have at least 2 tracks that contribute to the probabililty function. Again, the probability distribution is flat, with a spike at zero attributable to particles with a lifetime. The smaller spike at one is believed to result from events where the jet in question dominated the primary vertex fit, giving rise to a correlated set of tracks with artificially low impact parameters. If we attribute the excess of events in the first bin to a real heavy flavor content in this QCD sample and use the measured efficiency and background rates described below, we infer a b fraction in the QCD sample of approximately 5%. This number is in agreement with the fraction obtained by studying positive decay-length Jetvtx tags[3] in this same jet sample.

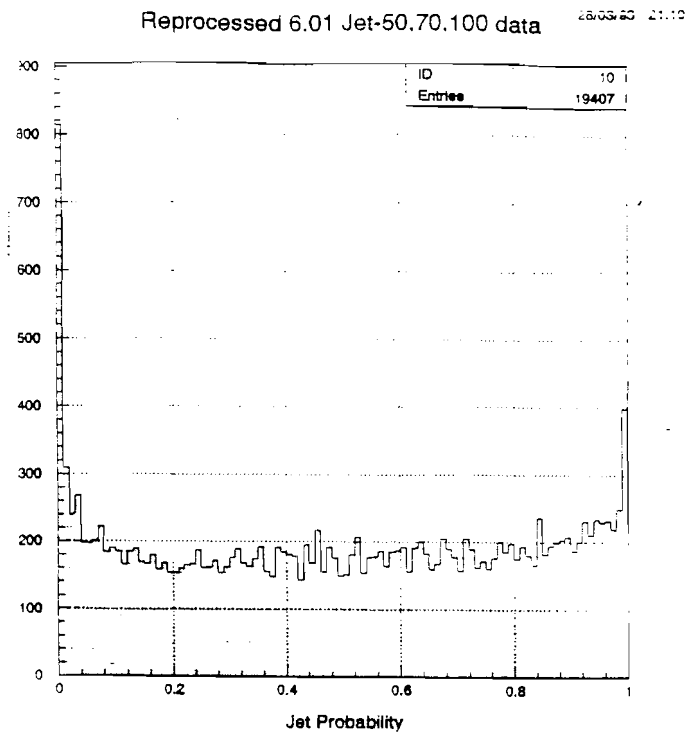

Figure 7: Jet probability distribution for jets in the QCD sample.

The jet probability distribution for muon jets in a $b \rightarrow \mu$ Monte Carlo sample is shown if Fig. 8. (The generation-level P_T cut on the b's in this sample is 15 GeV.) The jet probability distribution for the recoil jet in this sample is shown in Fig. 9. Recoil jets in b-$\bar{b}$ events are approximately 45% b's and 55% recoiling gluons, and this is reflected in the larger flat background compared to that in Fig. 8. Fig. 10 shows the jet probability distribution for recoil b jets in the same Monte Carlo sample; it closely resembles that of the leptonic b jet in Fig. 8. These plots suggest the strong ability of the jet probability tag to discriminate b-jets from non-b jets. We now turn to the data to measure the efficiency as a function of the jet probability cut.

3 Measurement of Efficiency

We have previously measured[6] the Jetvtx and D-ϕ b-tag efficiencies by studying tag rates in the inclusive 9 GeV CMU-CMP muon sample. We have used this same sample to measure the efficiency of the jet probability tag, and the numbers presented below may be compared directly to those in CDF 1962, which also gives a detailed discussion of the techniques used. Briefly, we measure the efficiency in three ways:

1. **Lepton Jet Tags.** The rate of tags in the muon jet is measured and normalized to the measured b fraction of 40%. The b fraction was measured[6] by fitting the signed impact parameter distribution of the muon to a superposition of b, c, and generic jet distribtions.

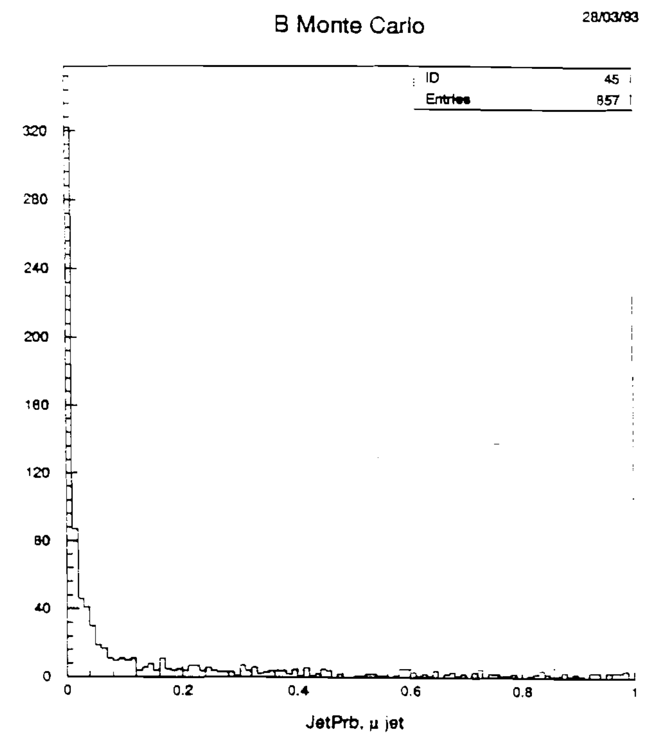

Figure 8: Jet probability distribution for the muon jet in a b-$\bar{b}$ Monte Carlo sample.

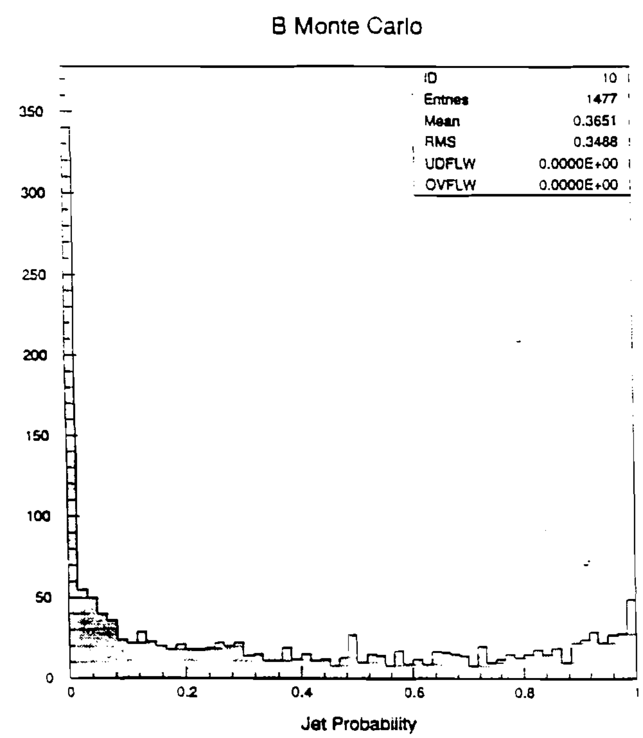

Figure 9: Jet probability distribution for the recoil jet in a b-$\bar{b}$ Monte Carlo sample. The recoil jets are approximately 45% b's.

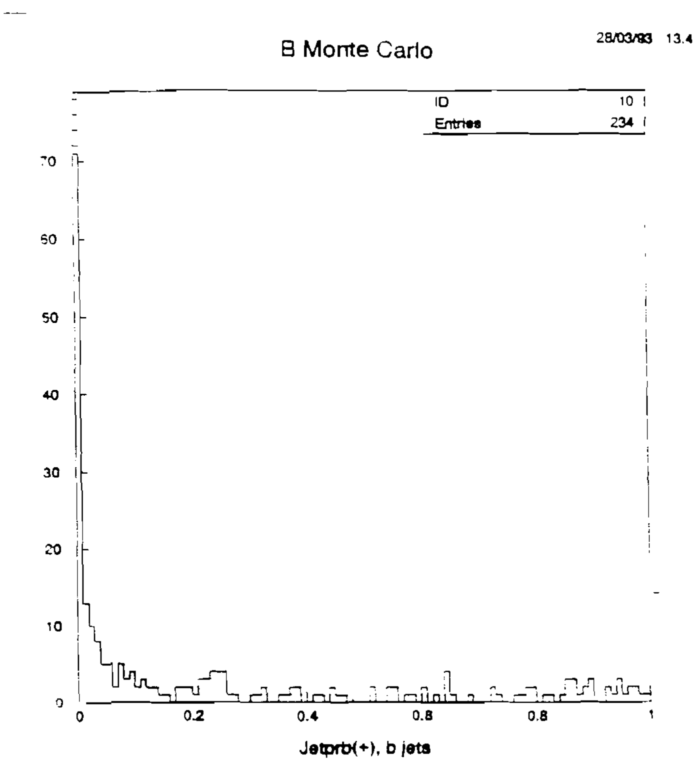

Figure 10: Jet probability distribution for the recoil b jet in the b-$\bar{b}$ Monte Carlo sample.

2. **Recoil Jet Tags.** The rate of tags in the recoil, or "away" jet, is measured and normalized to the measured b fraction times the fraction of recoil jets that are b's. This fraction is estimated from Monte Carlo calculations to be $45 \pm 15\%$. In fact, the jet probability distribution itself appears to offer a very promising method to *measure* this fraction with high precision, and work on this is in progress.[2]

3. **Double/Single Tag Ratio.** In this technique, a sample of events with tagged away jets provides an almost pure b-$\bar{b}$ sample. We then study the rate of tags in the muon jet in these events. This technique, while lower in statistics, does not require knowledge of the b fraction.

The efficiencies we report are the efficiencies per SVX-fiducial jet. An SVX-fiducial jet, in the b-tag group's agreed-upon definition, is a jet with at least two associated SVXS tracks (of any quality) in a cone of 0.4 centered on the jet axis.

3.1 Muon Jet Tags

Fig. 11 shows the muon jet probability distribution in the inclusive 9 GeV muon sample. It looks, as expected, like the distribution from the Monte Carlo in Fig. 8

[2]The prospect of tagging the away jet with high efficiency of course opens up a rich menu of b-physics topics, such as measurements of the correlated b-$\bar{b}$ cross section, the gluon splitting function, and the E_T-dependence of these quantities, to name just a few.

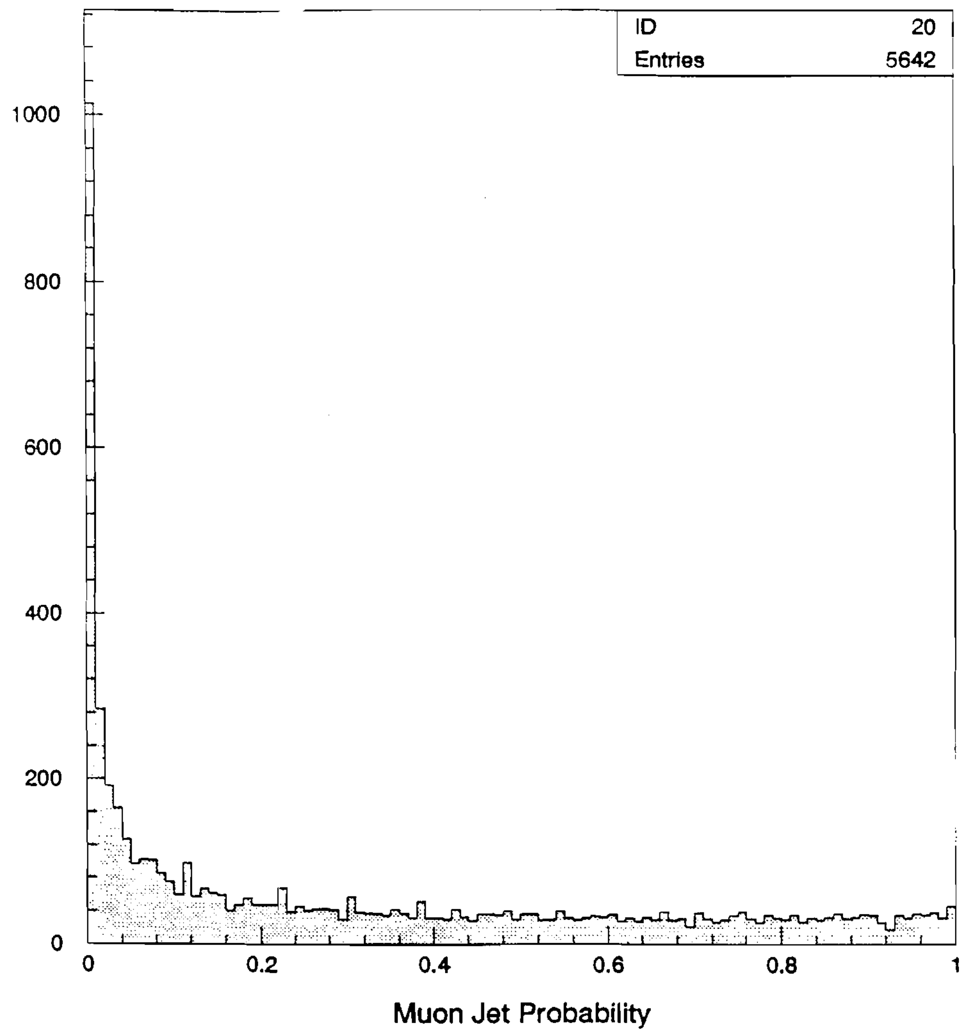

Figure 11: Jet probability distribution for the muon jet in the inclusive 9 GeV CMU-CMP muon sample.

with the addition of a flat background from the estimated 60% of this sample that is non-b. Normalizing to the b fraction and integrating the distribution, we find the efficiency per SVX-fiducial jet as a function of the jet probability cut. Because the efficiency is a rapidly-rising function for small values of this cut, we have plotted the efficiency vs. the logarithm of the jet probability cut in Fig. 12.

3.2 Recoil Jet Tags

Fig. 13 shows the jet probability distribution for the recoil jets in the same sample. Compared to Fig. 11, we see a larger flat background, confirming the expectation from Monte Carlo that fewer than half of the recoil jets in these b events are actually b-jets. Normalizing to the expected b fraction in these jets and integrating as before,

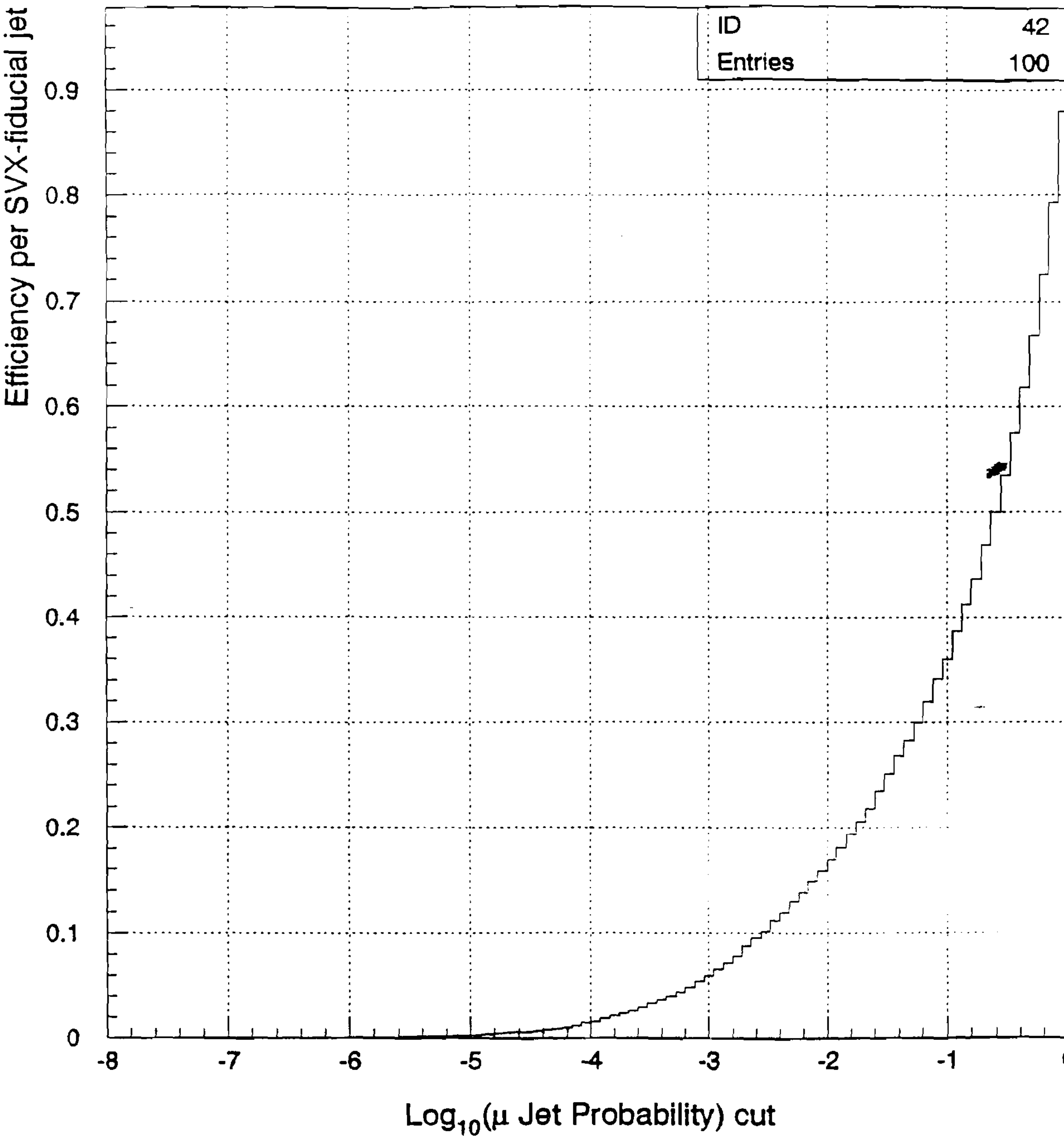

Figure 12: Efficiency of the jet probability tag as a function of the jet probability cut, as measured using muon jets.

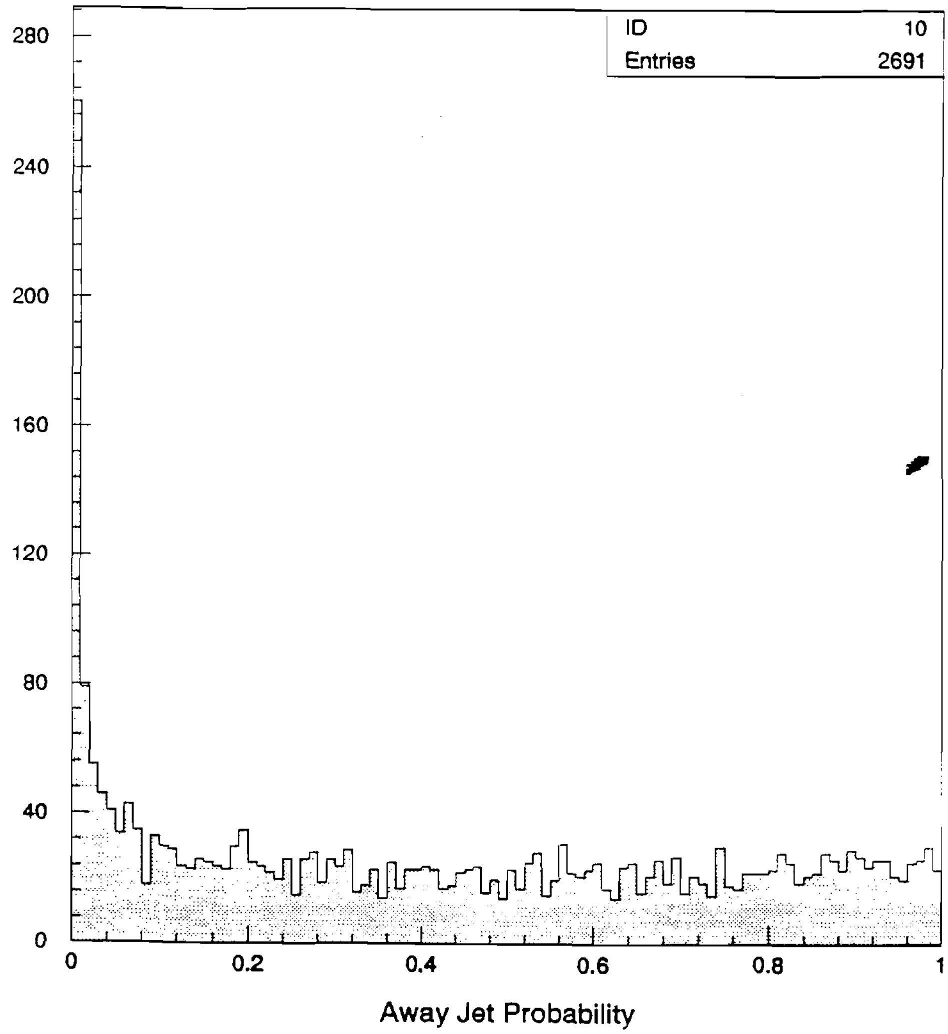

Figure 13: Jet probability distribution for the recoil jet in the inclusive 9 GeV CMU-CMP muon sample.

we obtain the efficiency curve shown in Fig. 14. In this curve, we have also subtracted a few-percent background estimated from the jet probability distribution in the QCD sample, Fig. 7. The backgrounds and normalizations were done in the region with jet probability less than 10%; points to the right of this in Fig. 14 are not meaningful.

3.3 Double/Single Tag Ratio

The most robust measurement of the efficiency, given adequate statistics, is provided by the double/single tag ratio, since this ratio is independent of the b fraction. Fig. 15 shows the jet probability distribution for the muon jet for events in which the recoil jet has a probability of less than 0.01. We expect these muon jets to constitute a nearly pure b sample, and indeed the similarity between Fig. 15 and the distribution

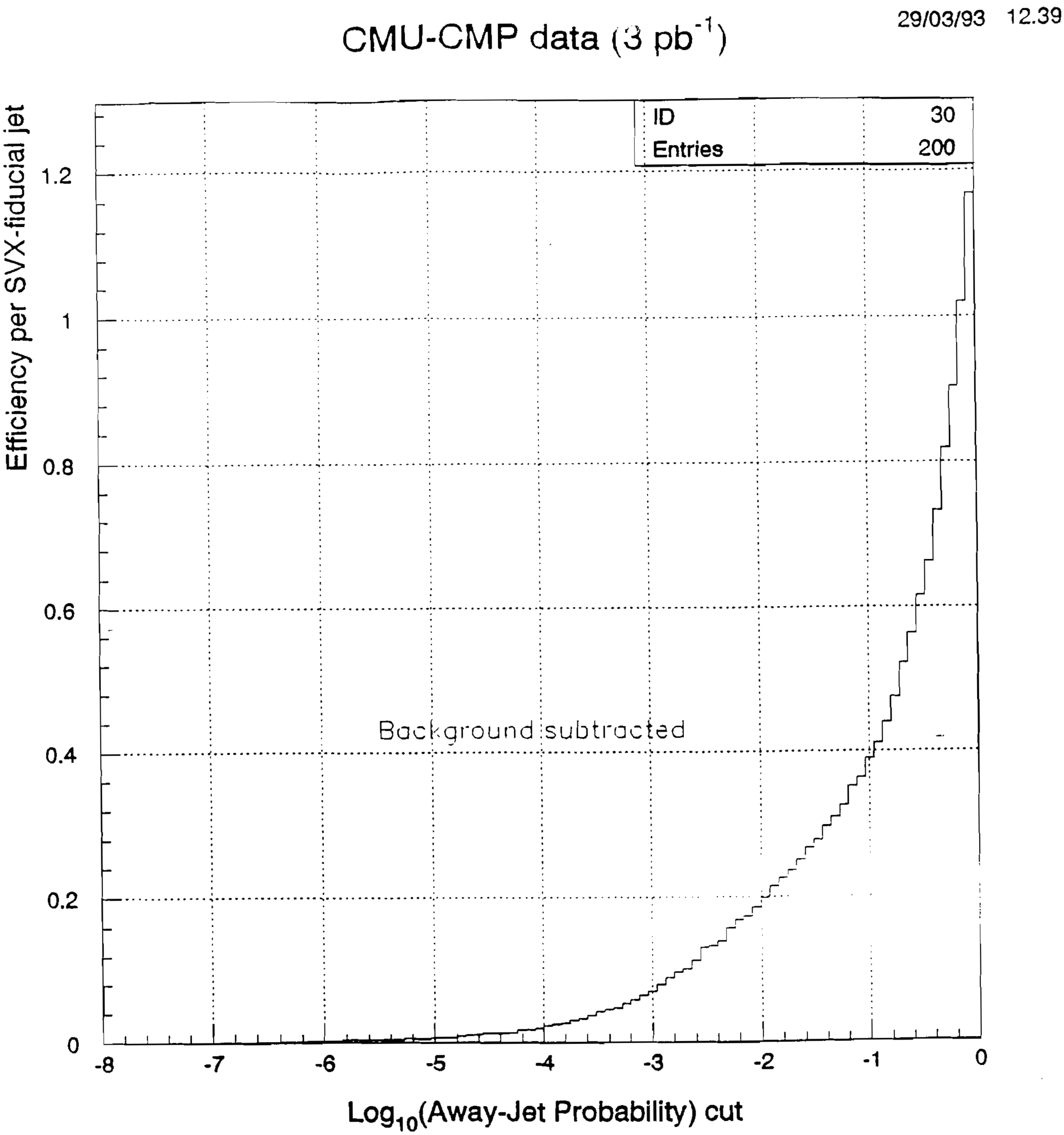

Figure 14: Efficiency of the jet probability tag as a function of the jet probability cut, as measured using recoil jets. In the region below 10%, the estimated background from generic jets has been subtracted.

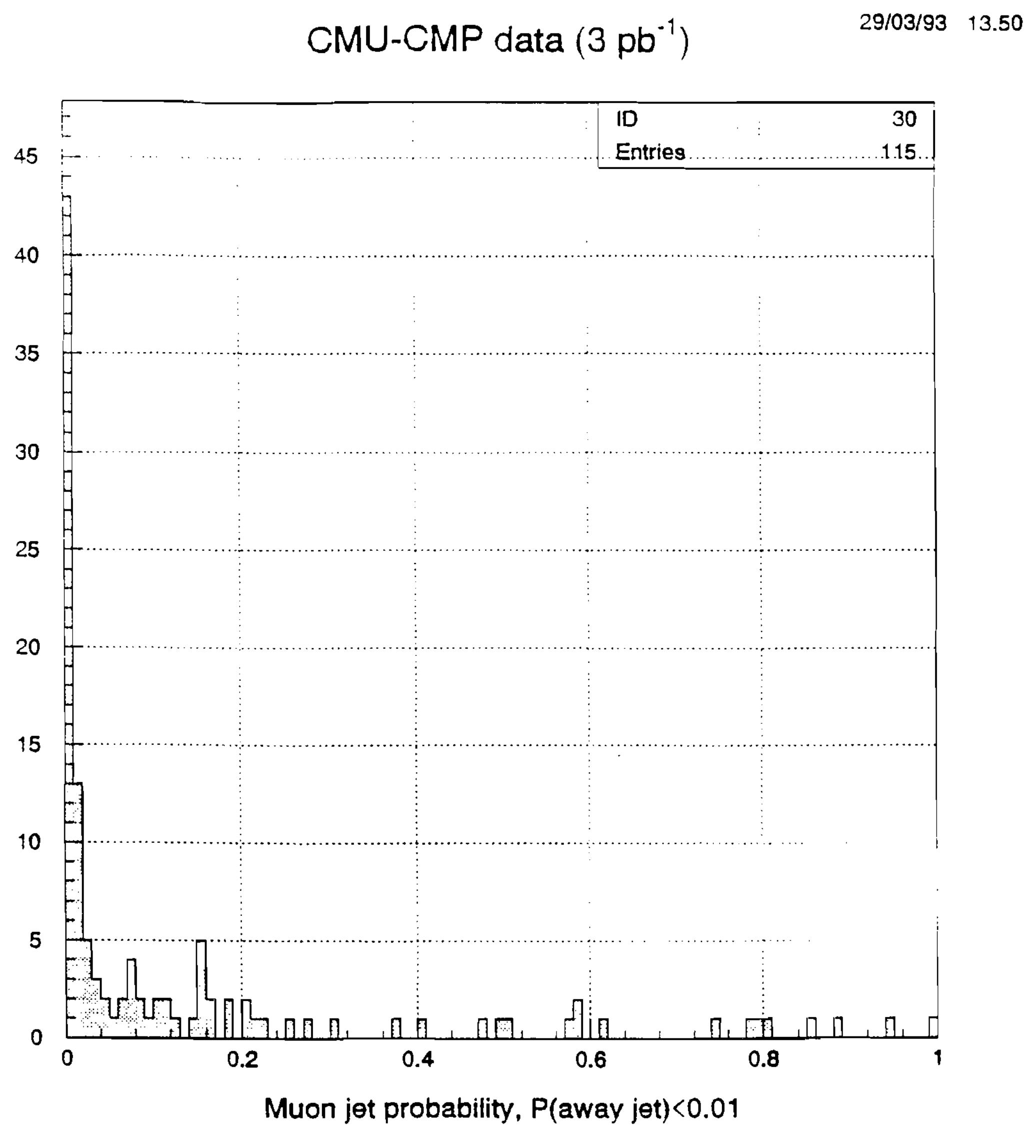

Figure 15: Jet probability distribution for the muon jet, in events where the recoil jet passed the $P_{jet} < 0.01$ tag.

from the b Monte Carlo (Fig. 8) is striking. Integrating this distribution, we obtain the efficiency curve shown in Fig. 16. Also shown, for reference, are the measured efficiencies for the Jetvtx, D-ϕ, and cone-tag algorithms.

All three techniques for measuring the jet probability tagging efficiency give consistent results, and all point to the prospect of achieving significantly higher efficiency than has so far been possible with other algorithms.

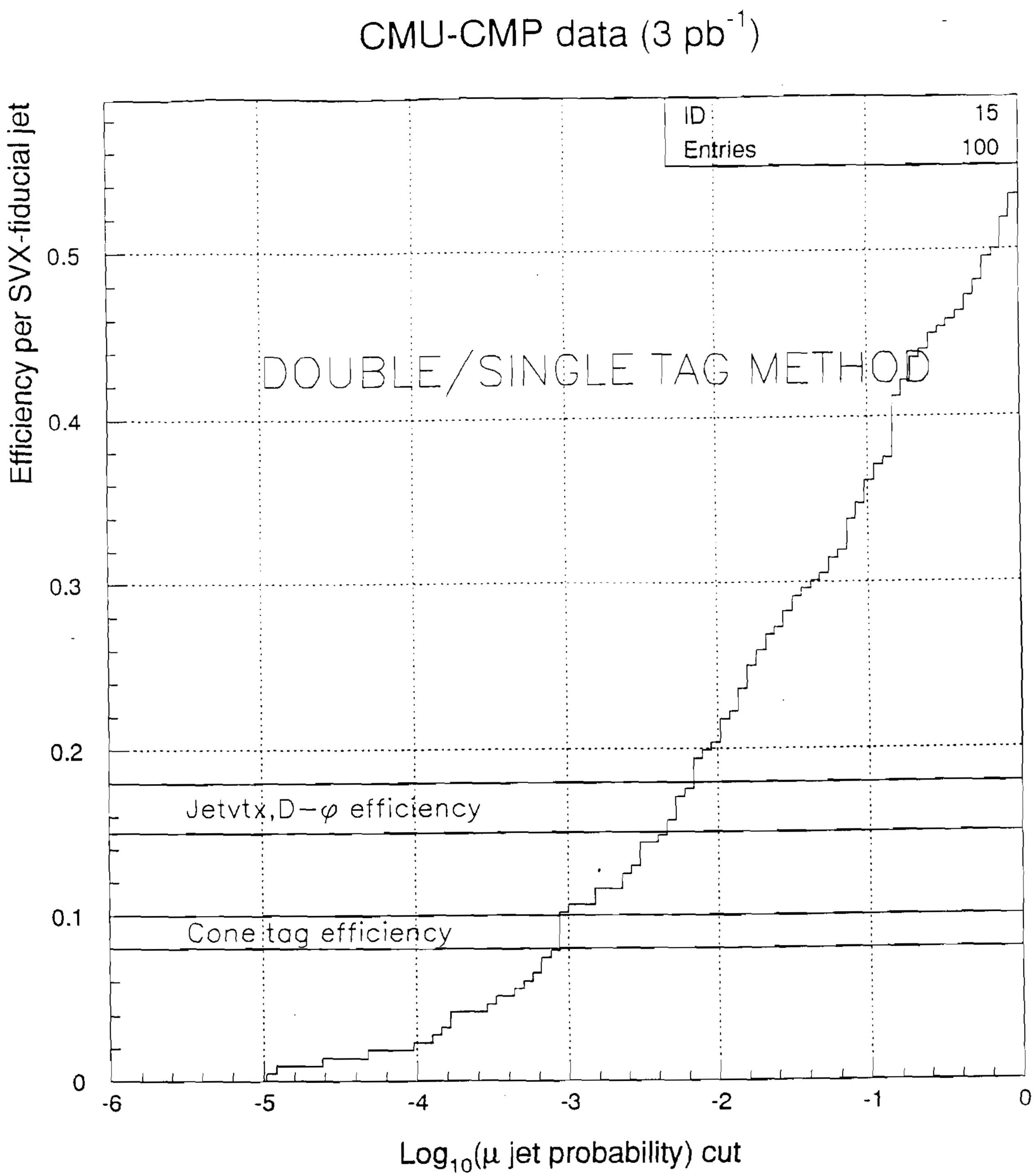

Figure 16: Efficiency of the jet probability tag as a function of the jet probability cut, as measured from the double/single tag ratio.

4 Background Rates

An attractive feature of the jet probability tag is its ability to provide many handles on background estimation. We obtain the background rate by studying tags in the QCD sample (the sample we used to measure the resolution function). Backgrounds may be divided into two categories:

1. Tags of no-life jets, caused by fluctuations consistent with the resolution function.

2. Tags of real heavy-flavor.

Because the probability distribution for no-life jets is flat, it can be easily estimated: if the probability cut for the b-tag is 0.01, the no-life background rate is 1%. A second handle on the no-life background rate is provided by the very useful control sample of *negative*-signed impact parameter tracks. One can form the jet probability instead using only the negative signed-i.p. tracks; the number of events "tagged" with this variable gives an indication of the pure mis-tag rate, as opposed to background rate from real heavy flavor. (The negative-i.p. tracks in jets form a control sample that is in many ways analogous to the negative decay-length Jetvtx tags, albeit with much higher statistics.) Fig. 17 shows the jet probability distribution formed from the negative-i.p. tracks only. This figure should be compared to Fig. 7, which shows the same distribution formed using the positive-i.p. tracks. Comparing the two figures, we again see the peak near one (of approximately the same size and shape), believed to result from events where the jet dominated the primary vertex fit, giving all the tracks a systematically small impact parameter significance. There is also a much smaller peak near zero. This peak probably results from two effects:

1. The presence of more tracks with large negative impact parameter than are accounted for in the fit to the resolution function (see Fig. 4).

2. The mis-signing of tracks coming from particles with a lifetime. This can happen in cases where the jet axis is a poor approximation to the b-direction, or from certain tertiary decays.

We are working to understand the relative importance of these two effects. There is some evidence from the Monte Carlo that tracks from b-decay are mis-signed about 10–15% of the time. In Figs. 18 and 19, we show on a log scale the jet probability distributions formed using the positive- and negative-signed i.p. tracks respectively. The difference between the two curves can be attributed to real tags.

It is instructive to calculate the background rate as a function of track multiplicity (that is, the number of tracks that were used in the probability function, not the total number of tracks in the jet), for different values of the jet probability cut. Figs. 20, 21 and 22 show the tag rate vs. track multiplicity for for both the positive- and

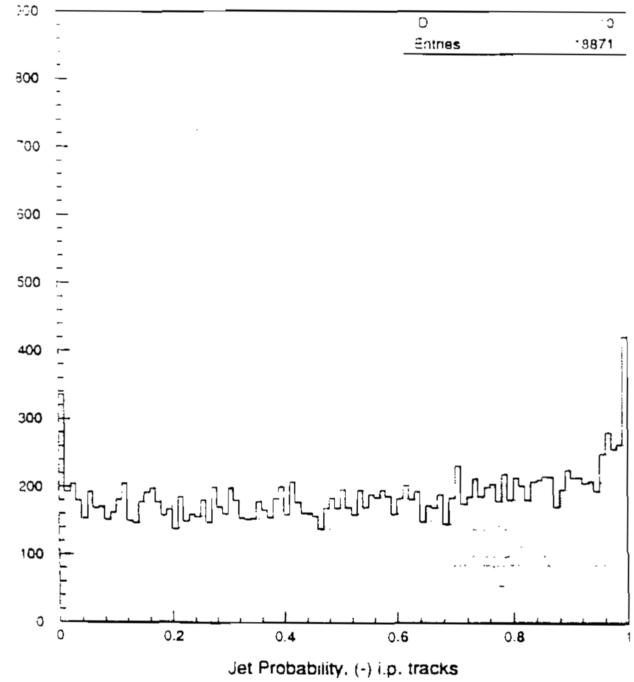

Figure 17: Jet probability distribution for jets in QCD sample, where the jet probability has been formed using the *negative* signed-i.p. tracks, rather than the positive ones.

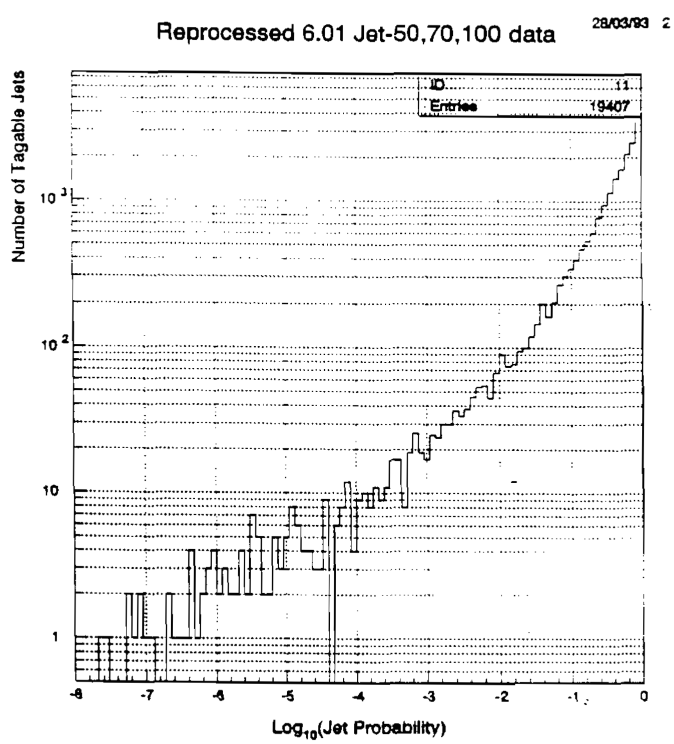

Figure 18: Jet probability distribtion for jets in the QCD sample, where the jet probability has been formed using only the positive-i.p. tracks

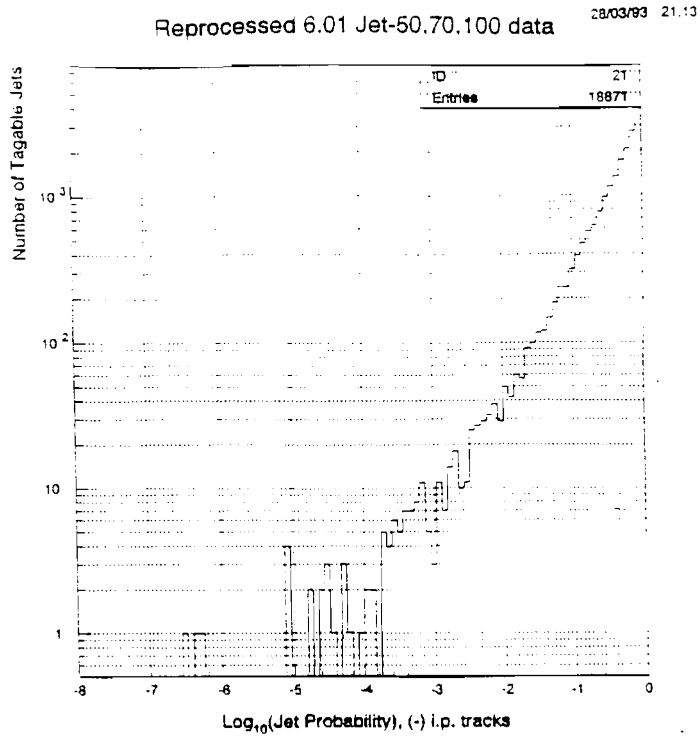

Figure 19: Same as the previous figure, but using the negative-i.p. tracks rather than the positives to form the jet probability.

negative-i.p. jet probabilities, for jet probability cuts of 0.001, 0.01, and 0.02. The b-tag efficiency in these units is shown on the figures. (The efficiency is higher that is shown in the efficiency curves in the last section—those were rates per SVX-fiducial jet, not all of which have 2 or more good tracks that can be used in the probability function. Here we are considering only "taggable" jets.) The rates rise with track multiplicity, and systematically rise and fall as the probability cut is changed. The ability to turn this "background knob" is an important and useful feature of the algorithm—in the top search, one can start, for example, with a loose probability cut and understand the backgrounds, then progresively tighten the cut and watch the backgrounds become small while the signal, hopefully, emerges in the 3- and 4-jet bins.

5 Conclusions

Many systematic issues remain to be studied, among them:

1. **Resolution function.** We have parametrized the resolution function as a function of the number of hits and number of shared hits on the track, but it could depend on other things. One obvious candidate is the P_T of the track, particularly at low-P_T where multiple scattering is most important. In addition, the resolution function may depend on the primary vertex error, and possibly on other things as well.

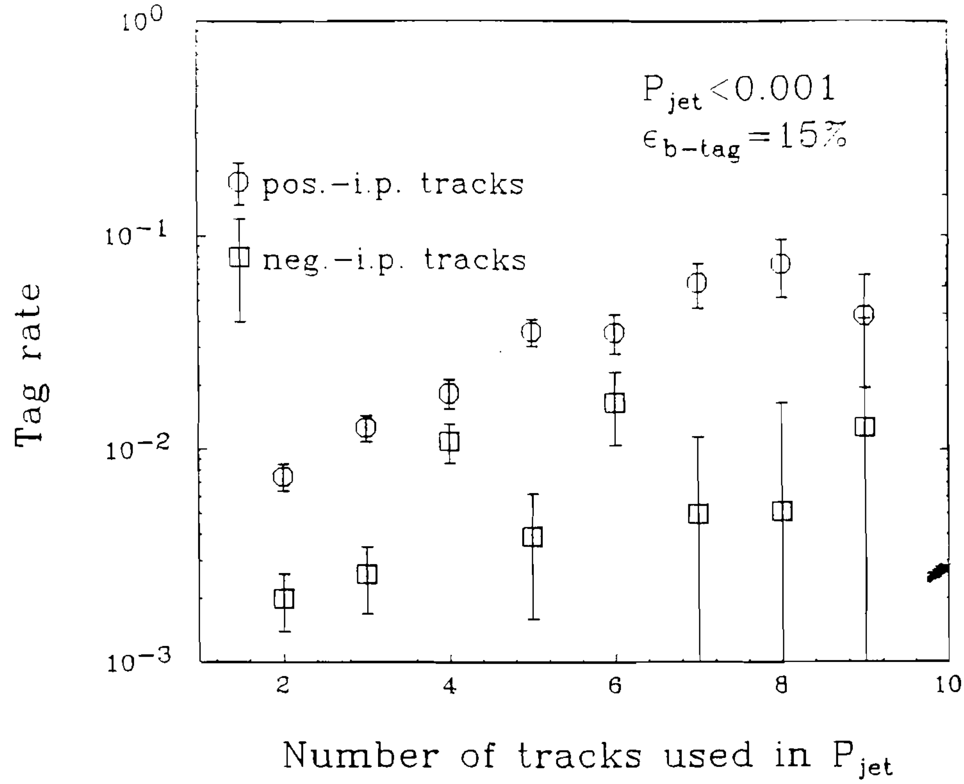

Figure 20: Tag rate in the QCD sample vs. track multiplicity, for the jet probability cut of 0.001.

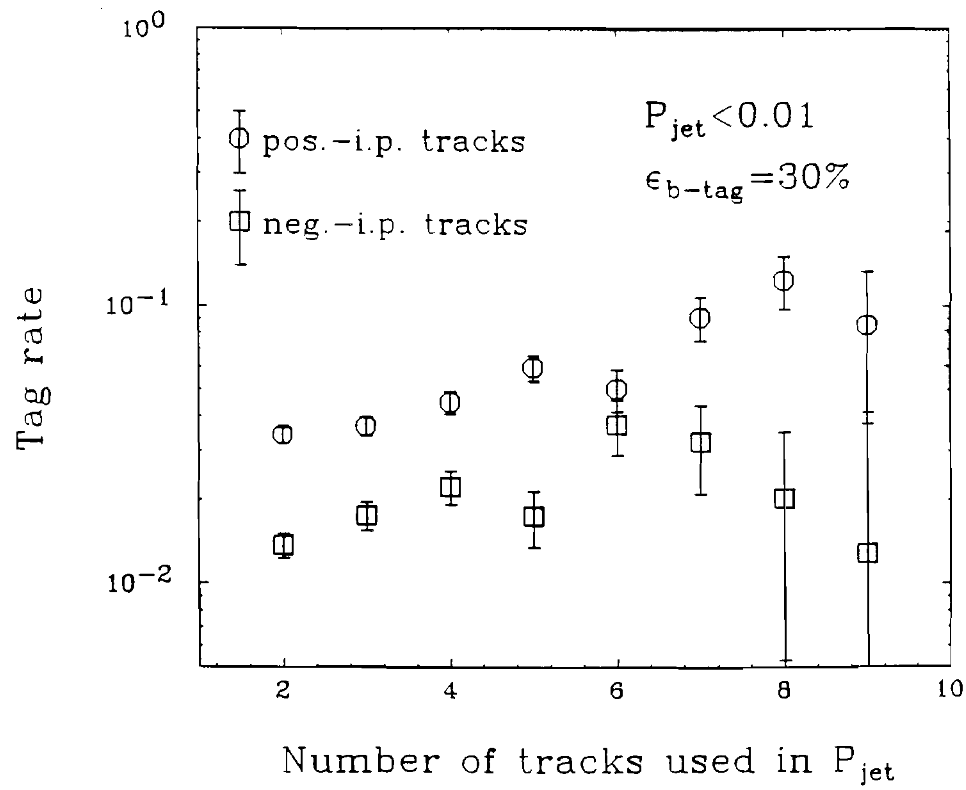

Figure 21: Tag rate in the QCD sample vs. track multiplicity, for the jet probability cut of 0.01.

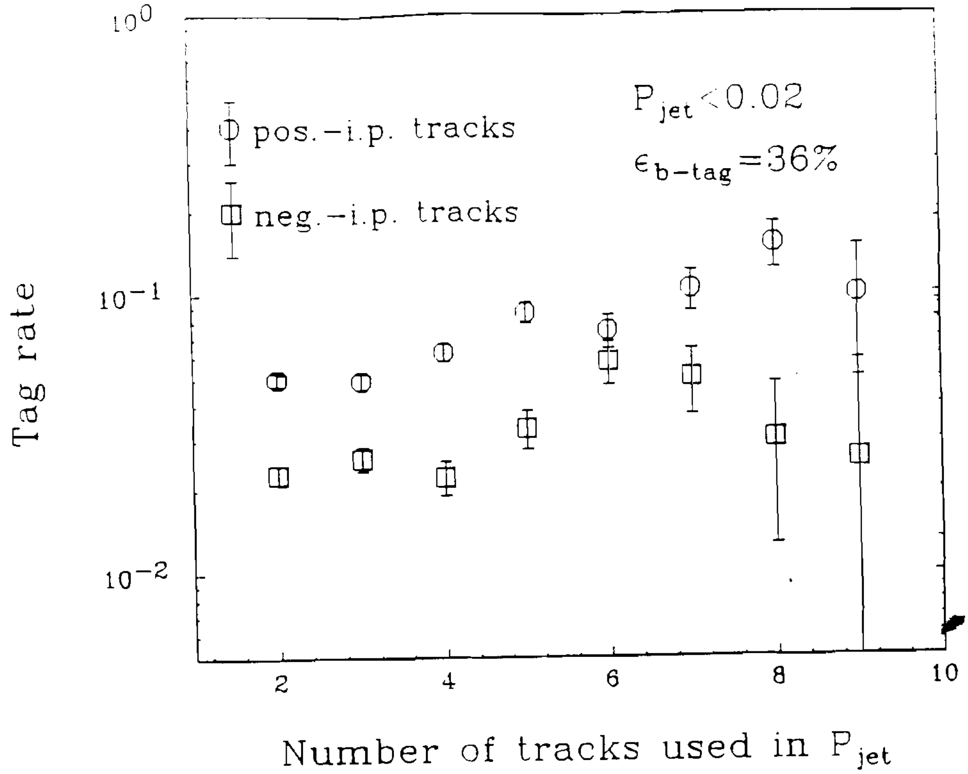

Figure 22: Tag rate in the QCD sample vs. track multiplicity, for the jet probability cut of 0.02.

2. **Primary vertex.** In addition to its possibly influence on the resolution function, we need to understandsuch issues as whether or not the primary vertex si better-determined in top events than in other events, and what effect this may have on our signal and background calculations. Is it better to use Vxprim or the beam position database?

Work continues.

Nonetheless, it is clear that jet probability tag is a powerful tool for b identification. The efficiency is higher than for other algorithms presently in use, and the backgrounds are low. Because the b-tag is imposed on a continuous variable, it is possible to optimize signal-to-background in a conceptually simple way. We expect many improvements in the top search with this technique.

References

[1] T. Chikamatsu *et al.*, CDF 1975.

[2] R. Hughes *et al.*, CDF 1959; C. Campagnari *et al.*, CDF 1961; D. Benton *et al.*, CDF 1967, CDF 2003; S. B. Kim *et al.*, CDF 2005.

[3] R. Hughes *et al.*, CDF 1959.

[4] G. Watts *et al.*, CDF 1956.

[5] S. Dell'Agnello *et al.*, CDF 1882.

[6] D. Amidei *et al.*, CDF 1962.

[7] D. Brown and M. Frank, Aleph Note 92-133.

[8] F. Bedeschi *et al.*, CDF 1939.