PROCEEDINGS OF SCIENCE

CDF *b*-tagging: Measuring Efficiency and False Positive Rate

Christopher Neu[*][†]
University of Pennsylvania
Philadelphia, PA 19104 USA
E-mail: neu@fnal.gov

The CDF experiment has developed several high p_T b-jet identification tools for the Run II physics program at the Tevatron. Herein we describe in detail one such b-tagging tool that exploits the long- lifetime of the b quark by identifying decay vertices significantly displaced from the primary interaction point. The b-tag efficiency is extracted from a b enriched data sample; the method is described, including a discussion of the important systematic effects. The data-driven measurement of the false positive tag rate is also described, as well as an explanation of how the per-jet false positive rate is used to predict the background contribution to the selected sample. Finally we conclude with a discussion of issues that have proven critical for b-tagging at CDF and should be given attention as we prepare b-tagging tools for LHC experiments.

International Workshop on Top Quark Physics
January 12-15, 2006
Coimbra, Portugal

[*] Speaker.

[†] On behalf of the CDF Collaboration.

1. Introduction

The ability to identify jets originating from b quark production, or b-tagging, is critical for several of the primary physics goals of the 2001-2009 run of the Tevatron $p\bar{p}$ collider at $\sqrt{s}$ =1.96 TeV. These goals include precision studies of the top quark, the search for the standard model Higgs boson, and many searches for particles from physics beyond the standard model. Top quarks are predicted to decay to a W and a b nearly 100% of the time. The most precise single measurement of the top quark mass [1] uses b-tagging to obtain a pure sample of well-reconstructed top quark candidates. For the M_H range $M_H < \sim 135$ GeV$/c^2$, the Higgs of the standard model is predicted to decay predominantly to b's; observation of the Higgs at the Tevatron in this particularly interesting mass range will depend on the quality of our b-tagging tools.

There are characteristics of b-jets that differentiate them from light flavor and charm jets:

- the long lifetime of the b quark
- the large mass of B hadrons
- the energetic semileptonic decay of B hadrons

Several b-tagging tools are available at CDF that attempt to exploit these distinguishing features of b jets. The focus here will be CDF's secondary vertex b-tagger, since it is the most widely used among CDF analyses. But each technique addresses b jet identification from a different approach, and hence it is possible that the combination of the techniques could provide additional tagging power.

1.1 The CDF Detector

CDF II is a general purpose detector designed to study the particles created in the $\sqrt{s}$ = 1.96 TeV proton-antiproton collisions provided by the Tevatron in Run II. A thorough description of the CDF detector can be found elsewhere [2]. A drawing of a quadrant of the longitudinal cross section of the CDF detector can be found in Figure 1.

Charged particle tracking is exceptional at CDF and plays a major role in the b-tag algorithms. Tracking is performed in four detector subsystems, each residing within a 1.4 T axial magnetic field provided by a superconducting solenoid. Three silicon tracking detectors provide tracking information out to $\mid \eta \mid < 2.0$. L00 resides directly on the beampipe, and is a single sided silicon detector designed to withstand the radiation environment typical of Run II luminosities. Immediately outside L00 in the region 1.5 cm $< r <$ 10 cm is the SVX II, a five-layer detector with axial and stereo silicon strips. The layers of the ISL are located at even larger radii. The stereo silicon strips provide three dimensional reconstruction of each track. The CDF silicon tracking provides impact parameter resolution of $\sim$ 40 μm, which includes the contribution from the beam that is approximately 30 μm wide.

The Central Outer Tracker (COT) is a cylindrical open cell wire chamber providing tracking information out to $\mid \eta \mid < 1.0$. The wires of the COT are arranged in eight layers, half of which are axial, and half are at a small stereo angle. On a given track there are up to 96 position measurements in the COT, and given an outer radius of 1.35 m, the COT has a large lever arm for curvature measurements. The COT provides track momentum measurements with resolution $\frac{\sigma_{p_T}}{p_T} = \sim 0.0015$ GeV $^{-1} \times p_T$.

In addition to charged particle tracking, CDF also contains electromagnetic and hadronic calorimetry up to $\mid \eta \mid < 3.6$. Muon detection is possible out to $\mid \eta \mid < 1.0$, although in the *b*-tagging studies described below muons from the most central portion of CDF will be considered ($\mid \eta \mid < 0.6$).

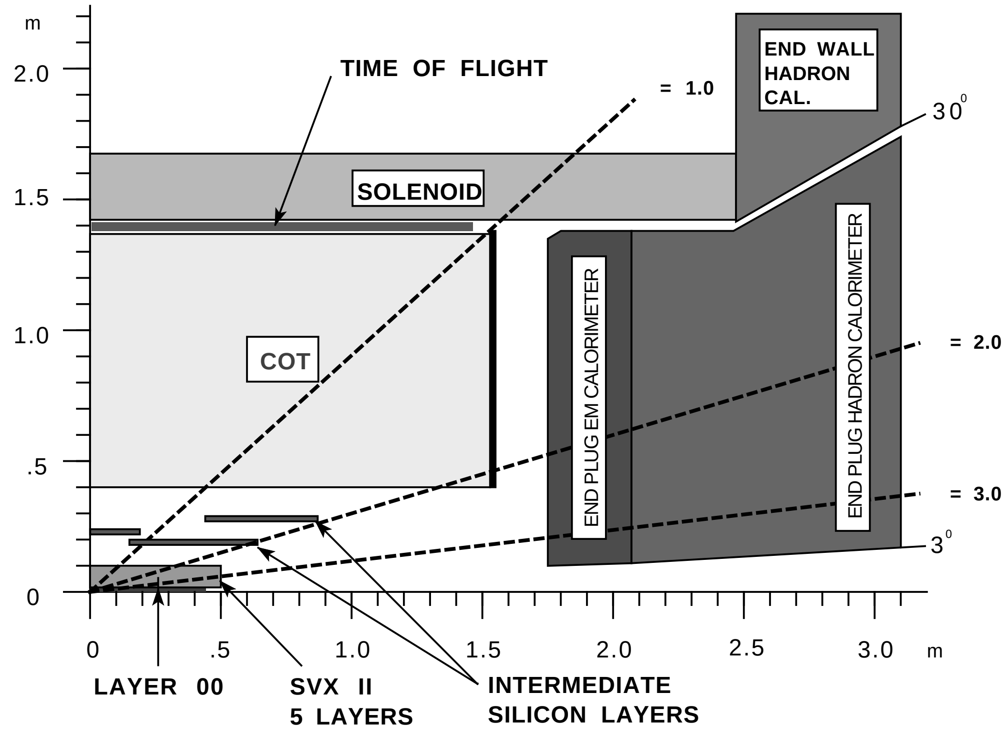

Figure 1: A cross-sectional quadrant view of the CDF detector focusing on the tracking systems.

2. Secondary Vertex Identification

The long-lifetime of *B* hadrons can be exploited in order to identify jets consistent with originating from *b*-quark production. Consider a *b*-quark produced in the decay of a top quark or Higgs boson at the initial p-$\overline{p}$ interaction point. The *b*-quark hadronizes almost immediately (on the order of 10^{-24}s) to form a jet of particles; included in this jet are a *B* meson (B^0, $B^{\pm}$,B_s^0) or a *B* baryon (Λ_B). The *B* hadron usually carries off most of the original *b*-quark momentum and has a relatively long lifetime (of order several ps). Given their long lifetime and large boost, *B* hadrons created in this way travel a macroscopic distance away from the primary interaction point in the lab frame before decaying into several charged and neutral particles. Reconstruction of charged particle tracks enables us to look for the trajectories of decay daughters that are inconsistent with originating from the initial interaction point. Several of these tracks can be determined to originate from a common location, and a so-called secondary vertex can be constructed.

The CDF secondary vertex detection algorithm, SECVTX, is designed to examine the tracks with large impact parameter (d_0) within each jet and to attempt to vertex them to a common point. Each track's impact parameter is measured with respect to a primary $p\overline{p}$ interaction position, or primary vertex, that is determined for each event. The SECVTX algorithm runs on a per-jet basis within each event. The algorithm starts by considering silicon tracks within each jet ($\Delta R <$ 0.4, where $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$). The silicon tracks must be seeded by or confirmed by a track in the COT. To be considered for SECVTX, the tracks within the jet are demanded to have $p_T >$ 0.5 GeV$/c$, d_0 significance $S_{d_0} \equiv |\frac{d_0}{\delta_{d_0}}| > 2.0$ with respect to the primary vertex, and a minimum number of hits in the silicon tracking detectors. The hit requirements are a function of the detector geometries and the track reconstruction quality. Tracks are further demanded to not exceed a maximum d_0 requirement in order to protect against poorly reconstructed tracks as well as tracks from long-lived light flavor hadrons or nuclear interactions in the detector material.

The selected tracks are then ordered in p_T, and a secondary vertex is sought among these tracks. The construction of a 2-track "seed" vertex is first attempted among the qualifying tracks. If a seed vertex is found, the remaining tracks are considered for vertexing with the seed tracks. Each additional track is considered singly; after attaching all qualifying tracks to the vertex, the vertex χ^2 is recalculated, and tracks are iteratively pruned from the vertex if they contribute too greatly to the overall χ^2.

The 2D decay length of the fitted vertex with respect to the primary is defined as L_{xy}. If the pruned vertex retains three or more tracks, this vertex is then subject to a final round of quality cuts, including removal of vertices from material and nuclear interactions, as well as those consistent with the decay K_s and Λ, two prominent long-lived light flavor hadrons. Finally the vertex is demanded to have $S_{L_{xy}} > 7.5$, where $S_{L_{xy}}$ is the 2D decay length significance, defined as $S_{L_{xy}} \equiv |\frac{L_{xy}}{\delta_{L_{xy}}}|$. If the vertex satisfies all of the above criteria, a secondary vertex is defined to have been found.

If no candidate vertex is found, a second pass at vertex construction is made. Efficiency is gained by only requiring two or more tracks satisfying more stringent track quality requirements.

If a secondary vertex is found with either pass, the jet is said to be "tagged". If the dot product of the 2D displacement vector from the primary vertex to the secondary vertex and the jet's momentum vector is positive (*ie.*, the vertex is in the same hemisphere of the detector as the jet), the tag is called "positive". If the secondary vertex and jet momentum have a negative dot product (the vertex is in the opposite hemisphere of the detector actually behind the jet), the tag is called "negative". Such vertices cannot be consistent with heavy flavor decays and are due to the finite tracking resolution of the CDF detector.

CDF supports three operating points for the SECVTX algorithm with different values of efficiency and purity: Ultratight, Tight, and Loose. The results discussed in the rest of the article are for the Tight operating point as this is most widely used in CDF analyses. The Loose algorithm allows increased efficiency for analyses that require two or three b-tagged jets. The Ultratight algorithm allows increased purity for analyses that require a single b-tagged jet.

As with any particle identification technique, when considering a b-tagger like SECVTX it is important to understand how often one tags a b jet in the data and how many of the tagged jets actually come from b's. The issues of efficiency and purity are discussed below.

2.1 Tag Efficiency

The efficiency of the SECVTX algorithm is defined as the fraction of b jets fiducial to the CDF COT and calorimetry that possess a positive SECVTX b-tag. Measuring the efficiency of a b-tagging algorithm is straightforward in Monte Carlo events; one has the luxury of complete knowledge of the particles within each jet, and thus it is straightforward to identify the fiducial jets that come from b quark production and the fraction which are tagged. But this is not to say that the efficiency measurement in Monte Carlo jets is accurate. Reliable modeling of b-tagging in the Monte Carlo requires precise understanding of the charge deposition in the silicon detectors, accurate simulation of the tracking, and realistic B hadron production and decay models. Since none of these effects are perfectly modeled in the Monte Carlo, it is imperative to measure the b-tag efficiency in the data.

The challenge in measuring the tag efficiency in data events is that the nature of individual jets is not explicitly known. The tag efficiency measurement in data at CDF relies upon constructing a pure sample of b jets within the large dijet sample. Two methods currently in use at CDF utilize high p_T leptons matched to jets to identify jet pairs consistent with heavy flavor.

The first technique requires a high p_T muon to be buried within a jet. his so-called “muon-jet” is paired with a back-to-back jet, known as the “away-jet”,which is demanded to possess a positive SECVTX tag. This jet pair (one jet containing a secondary vertex, the other having evidence for a high p_T semileptonic decay) is consistent with coming from heavy flavor production. By further requiring that the effective mass of the tracks in the secondary vertex of the away-jet be large ($M_{vtx} > 1.5$ GeV$/c^2$) to reduce the contribution from $q\bar{q}$ and $c\bar{c}$, the dijet sample is enriched in b jets. These criteria construct a muon-jet sample that is $\sim$ 78% pure with b’s.

The muon-jet provides the sample in which the efficiency measurement is completed. The p_T of the muon relative to the jet axis (p_T^{rel}) is a powerful discriminator of b jets from jets from charm and light flavor. One can construct p_T^{rel} templates for b and non-b jets from Monte Carlo dijet events and then fit for the number of b jets in the tagged and untagged samples (see Figure 2). From the b fractions, one can calculate the efficiency for SECVTX to tag b jets in the data. With this technique, the tag efficiency in 350 pb^{-1} of Run II CDF data is measured to be 0.39 $\pm$0.01(stat only), integrated over the complete jet E_T range. The SECVTX tag efficiency depends strongly on jet kinematics (E_T and η); these jet properties will clearly vary depending on the physics process one considers. These effects should be considered when determining the tag efficiency in the context of an individual measurement.

The tag efficiency is used most often when assessing signal acceptance, which is typically done in signal Monte Carlo samples. As discussed above, Monte Carlo b-jets are not guaranteed to perfectly match data b jets. So it is necessary to construct a data-to-Monte Carlo scale factor for tag efficiency, which encapsulates the discrepancies between b-jet tagging in the Monte Carlo and data. In an appropriate dijet Monte Carlo sample matching the conditions used above to select the muon-jet sample, the tag efficiency was measured to be 0.43 $\pm$0.002 (stat only). The scale factor then is calculated to be 0.92 $\pm$ 0.02(stat) $\pm$ 0.06(syst). The major source of systematic error is the jet E_T dependence of the measurement: the muon p_T^{rel} method utilizes jets with small E_T values (20-40 GeV), while most of the interesting physics we seek to do with high p_T b tagging exists at larger E_T values (50-70 GeV in top quark decay), and extrapolating scale factor results to the large

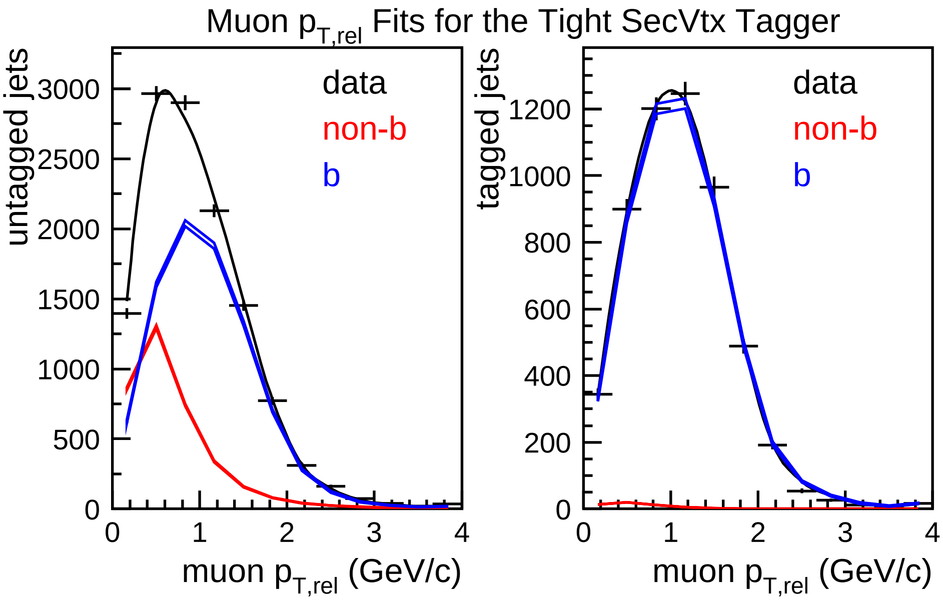

Figure 2: Fits of the untagged and tagged data muon-jet sample p_T^{rel} distributions using b and non-b templates from MC. Jets from non-b sources have typically smaller p_T^{rel} values, as one can see from the fit to the untagged muon-jet sample, where the contribution from non-b jets is significant.

E_T regime suffers from lack of statistical power. Additional systematic errors include the imperfect detector response in the jet energy measurement and the reliance on a sample of b jets containing semi-leptonic decays and applying to non-semileptonic decaying jets. The value of the scale factor being less than unity indicates that the tagging conditions in the Monte Carlo are more optimistic than reality.

The second tag efficiency measurement method employed at CDF is similar to the muon p_T^{rel} measurement but uses electrons matched to jets to achieve a highly enriched b sample. Similar demands are placed on the away-jet to increase the b purity. The subsample of the electron-jets that have a photon conversion partner are separated from the remainder of the jets; since electrons from semi-leptonic hadron decay should not be consistent with coming from a conversion, these jets provide a complementary sample with reduced b purity which is used to determine the light flavor content of the away-tagged sample. The electron-jets inconsistent with photon conversions are the sample in which the data tag efficiency is measured. An algebraic solution is determined for the data tag efficiency, and then, given the simple MC tag efficiency, the tag scale factor from the electron method can be computed as well. The result is found to be 0.89 ± 0.03(stat) ± 0.07(syst), which is in good agreement with the result of the muon p_T^{rel} method using the same integrated luminosity.

Equipped with the data-to-Monte Carlo tag efficiency scale factor, one can then determine the tag efficiency for signal events in the data. For this purpose a sample of Monte Carlo $t\bar{t}$ events was produced in Pythia [3], and the tag efficiency was measured as a function of b jet E_T and $\mid \eta \mid$ in top decay, as shown in Figure 3. Efficiency curves are shown for the Tight and Loose SECVTX operating points. The efficiency measured in the Monte Carlo jets is multiplied by a combined scale factor derived from the two results discussed above, so the efficiency curves reflect the actual tag efficiency in the data. One can see that the tag efficiency reduces slightly as jet E_T increases; this is due in part to the collimation of tracks in high E_T jets and the difficulties vertexing such tracks together. The removal of vertices from material interaction also contributes to the decrease in efficiency at high jet E_T; in such jets it is possible for the reconstructed secondary vertex to be located beyond the beampipe radius ($r_{beampipe} = 1.26$ cm), and thus qualify for removal as a material interaction. The efficiency decrease at large jet E_T is less dramatic when considering Loose SECVTX, for which the removal of these vertices is not a part of the algorithm. The efficiency reduction for $\mid \eta \mid > 1.0$ is due to reduced tracking efficiency and silicon coverage in the forward region. Recall that silicon tracks considered in SECVTX are matched to a track in the COT. Large $\mid \eta \mid$ tracks exit the COT through its endplate, thus reducing the number of layers traversed. This results in a decreased track efficiency in the region beyond $\mid \eta \mid = 1.0$. Currently development is proceeding on exploiting standalone silicon tracks in the forward region to increase *b*-tag efficiency.

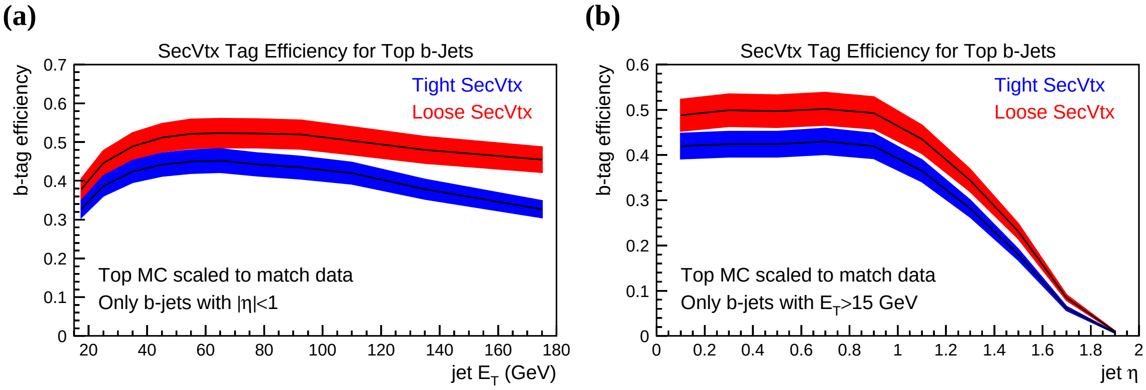

Figure 3: Tag efficiency for *b* jets in top decay in data as a function of jet E_T (a) and $\mid \eta \mid$ (b) for two SECVTX operating points.

The prospect of using actual $t\bar{t}$ events for *b*-tagging calibration at the LHC experiments is discussed elsewhere [4]. This is an attractive approach in that it directly provides information on the performance of the tagger in jets in the higher E_T regime. As noted above, one of the limitations of the standard efficiency measurement techniques employed at CDF is the extrapolation from the E_T range of the calibration sample to the E_T range of the signal sample; it is hoped that by utilizing jets in top events that the effect of this systematic error could be reduced or eliminated. Efforts towards this end are currently being pursued at CDF.

2.2 False Positive Rate

False positive tags, or mistags, in SECVTX come from the spurious identification of a secondary vertex in a non-*b* jet. Jets from light flavor production should be consistent with zero lifetime. However tracks within a light flavor jet can still have large impact parameter and hence satisfy the secondary vertex requirements. Sources of such spurious large impact parameter tracks include:

- limited detector resolution
- long-lived light particle decays (Λ,K_s)

- material interactions

Mistags due to limited detector resolution are expected to be symmetric in the signed 2D displacement L_{xy} of the vector separating the secondary and primary vertices. One can then use the ensemble of negatively tagged jets ($L_{xy} < 0$) as a prediction to the light flavor jet contribution to the positive tag sample.

At CDF an *a priori* prediction of the mistag rate is calculated from the inclusive jet samples. The inclusive jet sample is collected on a set of simple triggers that collect events with minimum amounts of calorimeter energy; four trigger samples are collected, with minimum jet E_T of 20, 50, 70 and 100 GeV. These samples are used for calibration of the mistag rate.

The probability for a given jet to be a mistag is determined from the probability that the jet is a negative tag. This probability comes from a per-jet negative tag parameterization in five variables:

- E_T
- ϕ
- η
- track multiplicity
- $\sum E_T^{jets}$

The parameterization is built from the inclusive jet samples and then is used to predict the number of mistags in the standard signal data samples. An example of a signal sample for $t\bar{t}$ production is the collection of events with a high p_T lepton and missing transverse energy, $\not{E}_T$ (indicative of $W^{\pm}$ decay), and several jets. The jets in such a sample are considered on a jet-by-jet basis, and the probability of each jet to be a negative tag is extracted from the five-variable parameterization. From the mistag probabilities from the complete sample of jets in the signal sample, one can then predict the contribution from mistagged positive jets. These predictions have been shown to be valid within an 8% systematic error, which is dominated by the choice of calibration sample used to make the mistag parameterization.

However not all mistags are from resolution effects alone. Simply assuming that all mistagged jets are symmetric about the origin in L_{xy} will lead to an underestimate in the true rate of false positive tags. The contribution from long-lived particle decays and material interactions to the SECVTX positive tag rate has been studied as well. These contributions to the light flavor tag sample are at strictly positive L_{xy} values, thus introducing a light flavor mistag asymmetry. The mistag contribution to the positive tag sample is measured in the 50 GeV inclusive jet sample discussed above. Templates of the pseudo-$c\tau \equiv L_{xy} \times \frac{M_{vtx}}{P_{T,vtx}}$ distribution for b, c and light flavors are constructed from dijet Monte Carlo samples, and the contribution from each source is derived from a three-component fit to the data shapes. Using this technique the mistag asymmetry (A) was measured to be 1.36 ± 0.23(syst). The asymmetry is used to scale the predicted mistag rate from negative tags, R^{-}_{pred}:

$$\frac{N^{+}_{light}}{N_{light}} = A \times R^{-}_{pred} = (1.36 \pm\ 0.23) \times R^{-}_{pred} \tag{2.1}$$

With this prescription, the predicted mistag rate better matches the true contribution to the positive tag sample from non-HF sources. It also takes into account the (albeit small) contribution to the negative tagged sample from real heavy flavor jets. The systematic on the asymmetry is driven by the uncertainty in the heavy flavor fraction in the sample in which it is applied.

With the mistag parameterization and asymmetry, one can examine the mistag rate in a generic jet sample as a function of jet E_T and $\mid \eta \mid$ as shown in Figure 4. More energetic jets have more energetic charged particle tracks that pass the p_T requirements for the SECVTX algorithm; this gives an increased combinatoric fake rate and hence the mistag rate grows with jet E_T.

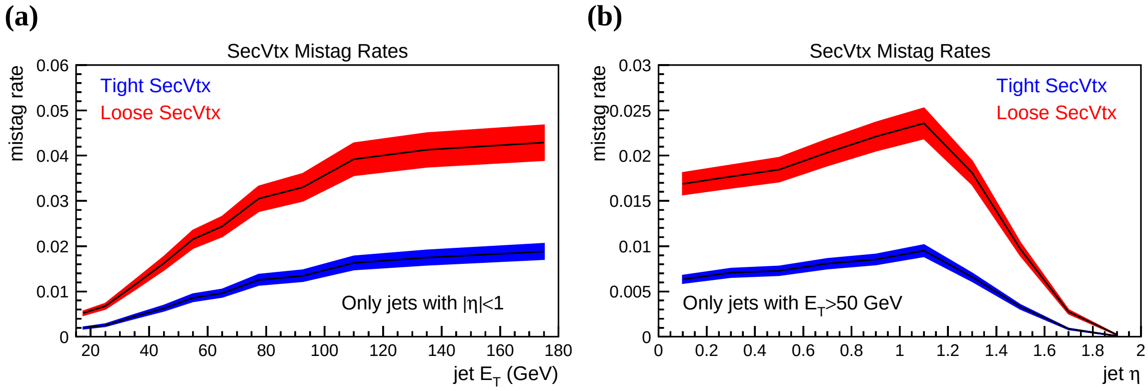

Figure 4: Mistag rate in data as a function of jet E_T (a) and $\mid \eta \mid$ (b).

3. Other *b*-tagging Strategies

The Soft Lepton Tagging algorithm [6] identifies *b*-jets by looking for evidence of a semi-leptonic *B* hadron decay within a jet. Semi-leptonic decay of hadrons is not unique to the *B* sector; however because of the large mass of typical *B* hadrons, the charged lepton to which the *B* decays typically has a higher transverse momentum with respect to the jet's axis than in non-*B* decays, and this feature is exploited when tagging jets with this technique. The CDF Soft Muon Tagger has been in use for several years; just recently development has begun on a Soft Electron Tagger.

JetProbability considers the d_0 of each track within a jet and constructs a probability that a given jet is consistent with coming from a zero-lifetime source. This probability distribution allows one to easily choose the efficiency/purity operating point by tuning the cut on the value of output probability. A virtue of JetProbability is that it provides a continuous output, which allows it to be used as a discriminant variable.

These disparate sources of tag information on jets lends itself to a combination, given that each tool approaches the task of *b*-tagging using different sources of information. At CDF recent efforts have focused on combining these tagging tools using a multivariate technique like a neural network. It is hoped that by exploiting all the tagging information simultaneously, the tag efficiency and purity of the neural network tagger will surpass those of the standard tools alone. The multivariate combination of taggers at CDF is still under development.

4. Conclusions and Implications for the LHC

Although we have shown techniques for understanding the performance of the CDF taggers, this does not mean that the challenges of *b*-tagging at a hadron collider are easy to overcome.

Future experiments, such as CMS and ATLAS at the LHC, should be mindful to pay close attention to several subtleties that are essential for quality *b*-tagging. For example, understanding the alignment of the tracking detectors is critical for precision measurement of track impact parameters, a necessary ingredient for lifetime-based *b*-tagging. The modeling of charge deposition for particles as they traverse the inner tracking detectors is important for high quality tracking, as is an accurate tracking simulation. The material content around the interaction region can have a significant impact on the tagging rate, and an effort should be made to understand the detector's true profile, including support structures, cables, cooling lines and readout electronics. And finally, well-understood *b*-tagging relies heavily on having access to calibration samples that as closely as possible reproduce the topologies of interesting signal events. Efficient triggering for events in the proper E_T and η range is therefore essential for both calibration and signal samples.

Much experience has been gained in the area of *b*-jet tagging at a hadron collider during Runs I and II at the Tevatron. Successful *b*-tag algorithms are currently in use at CDF. Here we have described in detail one of the CDF *b* tagging tools, the secondary vertex identification algorithm SECVTX. The techniques for measuring tag efficiency and mistag rate in the data have been described in detail. It is hoped that future experiments capitalize on this *b*-tagging expertise as we move forward into the energy frontier.

We thank the Fermilab staff and the technical staffs of the participating institutions for their vital contributions. This work was supported by the U.S. Department of Energy and National Science Foundation; the Italian Istituto Nazionale di Fisica Nucleare; the Ministry of Education, Culture, Sports, Science and Technology of Japan; the Natural Sciences and Engineering Research Council of Canada; the National Science Council of the Republic of China; the Swiss National Science Foundation; the A.P. Sloan Foundation; the Bundesministerium für Bildung und Forschung, Germany; the Korean Science and Engineering Foundation and the Korean Research Foundation; the Particle Physics and Astronomy Research Council and the Royal Society, UK; the Russian Foundation for Basic Research; the Comisión Interministerial de Ciencia y Tecnología, Spain; in part by the European Community's Human Potential Programme under contract HPRN-CT-2002-00292; and the Academy of Finland.

References

[1] A. Abulencia *et al*. [CDF Collaboration], *Precision top quark mass measurement in the lepton + jets topology in p anti-p collisions at s**(1/2) = 1.96-TeV*, Phys. Rev. Lett. **96**, 022004 (2006).

[2] D. Acosta, *et al*. [CDF Collaboration], *Measurement of the J/psi meson and b-hadron production cross sections*, Phys. Rev. D **71**, 032001 (2005).

[3] T. Sjostrand, S. Mrenna and P. Skands, *PYTHIA 6.4 physics and manual,* [hep-ph/0603175].

[4] See for example the talk by J. Heyninck of the CMS Collaboration at Top2006.

[5] D. Acosta, *et al*. [CDF Collaboration], *Measurement of the Cross Section for tt-bar Production in pp-bar Collisions using the Kinematics of Lepton + Jets Events*, Phys. Rev. D **71**, 052003 (2005).

[6] D. Acosta *et al*. [CDF Collaboration], *Measurement of the t anti-t production cross section in p anti-p collisions at s**(1/2) = 1.96-TeV using lepton plus jets events with semileptonic B decays to muons*, Phys. Rev. D **72**, 032002 (2005).